Smoothed Analysis of the Perceptron Algorithm for Linear Programming

Avrim Blum* John Dunagan†

Abstract

The smoothed complexity [1] of an algorithm is the expected running time of the algorithm on an arbitrary instance under a random perturbation. It was shown recently that the simplex algorithm has polynomial smoothed complexity. We show that a simple greedy algorithm for linear programming, the *Perceptron* algorithm, also has polynomial smoothed complexity, in a high probability sense: that is, the running time is polynomial with high probability over the random perturbation. While the bounds are not strictly comparable, for many choices of parameters the running time bound we show here for the Perceptron algorithm is much lower than the bound given in [1] for the Simplex algorithm.

1 Introduction

Spielman and Teng [1] recently proposed the smoothed complexity model as a hybrid between worst-case and average-case analysis of algorithms. The idea is to take an arbitrary (worst-case) linear program and then perturb each entry int he constraint matrix by a small random amount. They analyzed the running time of the simplex algorithm with the *shadow vertex* pivot rule for a linear program with m constraints in d dimensions, subject to a random Gaussian perturbation of variance σ^2. They showed that the expected number of iterations of the simplex algorithm was at most $f(m, d, \sigma)$, given as follows:

$$f(m,d,\sigma) = \begin{cases} \tilde{O}(\frac{d^{16}m^2}{\sigma}) & \text{if } d\sigma \geq 1, \\ \tilde{O}(\frac{d^5 m^2}{\sigma^{12}}) & \text{if } d\sigma < 1. \end{cases}$$

Each iteration of the simplex algorithm takes $O(md)$ time when we let arithmetic operations have unit cost. Spielman and Teng also speculate that their current analysis can be improved to yield an upper bound on the expected number of iterations of $\tilde{O}(\frac{d^5m^2}{\sigma^4})$.

*Department of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213. Supported in part by NSF grants CCR-9732705 and CCR-0105488. Email: `avrim@cs.cmu.edu`

†Department of Mathematics, MIT, Cambridge MA, 02139. Supported in part by NSF Career Award CCR-9875024. Email: `jdunagan@math.mit.edu`

In this paper, we show that a simple greedy linear programming algorithm known as the *perceptron* algorithm[2, 3], commonly used in machine learning, also has polynomial smoothed complexity (in a high probability sense). The problem being solved is identical to that considered by Spielman and Teng, except that we replace the objective function $\max c^T x$ by a constraint $c^T x \geq c_0$. In addition to simplicity, the perceptron algorithm has other beneficial features, such as resilience to certain types of random noise[4, 5, 6].

Specifically, we prove the following result, where all probability statements are with respect to the random Gaussian perturbation of variance σ^2. Note that each iteration of the perceptron algorithm takes $O(md)$ time, just like the simplex algorithm.

Theorem 1.1. (Perceptron Smoothed Complexity) *Let L be a linear program and let $\tilde{L}$ be the same linear program under a Gaussian perturbation of variance σ^2, where $\sigma^2 \leq 1/2d$. For any δ, with probability at least $1 - \delta$,*

either (i) the perceptron algorithm finds a feasible solution to $\tilde{L}$ in $O(\frac{d^3m^2\log^2(m/\delta)}{\sigma^2\delta^2})$ iterations

or (ii) $\tilde{L}$ is either infeasible or unbounded

The case of small σ is especially interesting because as σ decreases, we approach the worst-case complexity of a single instance. The theorem does not imply a bound on the *expected* running time of the perceptron algorithm (we cannot sample a new $\tilde{L}$ if we are unhappy with the current one), and thus the running time bounds given for the perceptron algorithm and simplex algorithm are not strictly comparable. Throughout the paper we will assume that $\sigma^2 \leq 1/2d$.

The perceptron algorithm solves linear programming feasibility problems and does not take in an objective function. However, given an objective function $\max c^T x$, we can use binary search on c_0 to find an approximately optimal $x \in \tilde{L}$. Specifically, suppose for a given linear program L and objective value c_0 we define $p_L(c_0)$ to be the probability (over the Gaussian perturbation) that there exists $x \in \tilde{L}$ with $c^T x \geq c_0$ and furthermore that $\tilde{L}$ is bounded. Then, the probability that the algorithm finds $x \in \tilde{L}$ such that $c^T x \geq c_0$ in $\tilde{O}(\frac{d^3 m^2}{\sigma^2 \delta^2})$ iterations (times the overhead of binary search on c_0) is at least $p_L(c_0) - \delta$. Since a solution with objective value c_0 or more only exists with probability $p_L(c_0)$ (unless $\tilde{L}$ is unbounded), this is a strong guarantee for the algorithm to provide.

The guarantee of theorem 1.1 is weaker than that of Spielman and Teng[1] in two ways. First, the simplex algorithm both detects and distinguishes between unbounded and infeasible perturbed linear programs, while we do not show a similar guarantee for the perceptron algorithm. Secondly, the simplex algorithm solves the perturbed linear program to optimality, while we show that the perceptron algorithm finds a solution which is good with respect to the distribution from which $\tilde{L}$ is drawn, but which may not be optimal for $\tilde{L}$ itself.

The high level idea of our paper begins with the observation, well-known in the machine learning literature (see, e.g., Minsky and Papert [7]), that the perceptron algorithm quickly finds a feasible point when there is substantial "wiggle room" available for a solution. We show that under random perturbation, with good probability, either the feasible set has substantial wiggle room, or else the feasible set is empty. In the remainder of the paper, we define the model of a perturbed linear program exactly (section 2), define the perceptron algorithm and prove a convergence criterion for it (section 3), transform the linear programming problem into the perceptron setting (section 4), state two geometric arguments (section 5), and finally prove our main theorem (section 6). We then give a short discussion of the meaning of our work in section 7. The proofs of several technical results are deferred to the appendices.

2 The Perturbation Model

We begin by restating the model of [1]. Let the linear program L be given by

$$
\begin{aligned}
&(2.1) & & \max \quad c^T x \\
&(2.2) & \text{s.t.} \quad & a_i^T x \quad \leq \quad b_i \quad \forall i \in \{1, \dots, m\} \\
&(2.3) & & |a_i| \quad \leq \quad 1 \quad \forall i \\
&(2.4) & & b_i \quad \in \quad \{\pm 1\} \quad \forall i
\end{aligned}
$$

As remarked there[1], any linear program can be transformed in an elementary way into this formulation. Now let $\tilde{a}_i = a_i + \sigma g_i$, where each g_i is chosen independently according to a d-dimensional Gaussian distribution of unit variance and zero mean. Then our new linear program, $\tilde{L}$, is given by

$$
\begin{aligned}
&(2.5) & & \max \quad c^T x \\
&(2.6) & \text{s.t.} \quad & \tilde{a}_i^T x \quad \leq \quad b_i \quad \forall i
\end{aligned}
$$

For completeness, we recall that a d-dimensional Gaussian is defined by the probability density function

$$\mu(x) = \left(1/\sqrt{2\pi}\right)^d e^{-|x|^2/2}.$$

3 The Perceptron Algorithm

The Perceptron algorithm is a classic algorithm for solving the following machine learning problem:

> 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 (w, w_0) such that for all positive points a_i, we have $a_i^T w > w_0$ and for all negative points a_i, we have $a_i^T w < w_0$.

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 a_i we rewrite the requirement $a_i^T w > w_0$ as $(a_i, 1)^T (w, -w_0) > 0$, 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 $a_i^T w < 0$ with $(-a_i)^T w > 0$.) The problem is now reduced to:

> Given a set of points $a_1, a_2, \dots, a_m$, find a vector w such that $a_i^T w > 0$ for all i, if one exists.

The Perceptron algorithm now works as follows:

Perceptron Algorithm

1. Initialize $w = \mathbf{0}$ (the all-zero vector).
2. Pick some a_i such that $a_i^T w \leq 0$ and update w by
$$w \quad \leftarrow \quad w + \frac{a_i}{|a_i|}$$
3. If we do not have $a_i^T w > 0$ 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 [7]) upper-bounds the running time in terms of the "wiggle room" available for a solution. Specifically, let w^* denote the solution that maximizes the wiggle room ν defined as $\nu =$

$\min_i \frac{|a_i^T w^*|}{|a_i||w^*|}$. In other words, not only is w^* feasible ($a_i^T w^* > 0 \quad \forall i$), but every w within angle $\arcsin(\nu)$ of w^* is also feasible.

THEOREM 3.1. (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. For completeness, we provide a proof of the theorem here.

Proof. ([7]) Let w^* be a solution of wiggle room ν, and for convenience scale 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 $w^T w^*$ and $|w|$. Notice that $w^T w^* \leq |w|$ since w^* is a unit vector. In each step, $w^T w^*$ increases by at least ν since $(w + \frac{a_i}{|a_i|})^T w^* = w^T w^* + \frac{a_i^T w^*}{|a_i|} \geq w^T w^* + \nu$. However, $|w|^2$ never increases by more than 1 in a given step since $(w + \frac{a_i}{|a_i|})^2 = w^2 + 2\frac{a_i^T}{|a_i|}w + (\frac{a_i}{|a_i|})^2 \leq (w^2 + 1)$, where we observed that $\frac{a_i^T}{|a_i|}w < 0$ for any i we would use in an update step. Therefore, after t steps we have $w^T w^* \geq \nu t$ and $|w| \leq \sqrt{t}$. The observation that $w^T w^* \leq |w|$ implies that $t\nu \leq \sqrt{t}$, 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 LP setting, the problem instance consists of hyperplanes (constraints) and the solution is a point. To transform our LP into the learning setting, we define the *polar* of a conic linear program.

Definition of Polar. For any d-dimensional space S filled with points and $(d-1)$-dimensional hyperplanes through the origin, we define the polar of S to be the d-dimensional space $P(S)$, where, for every point p in S, we define a hyperplane $p^T x = 0$ in $P(S)$, and for every hyperplane $h^T x = 0$ in S, we define a point h in $P(S)$. Because the feasibility problem $h^T x < 0$ is a cone, any feasible point 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.

4 Transforming the problem

We now show how to transform the linear program (2.5, 2.6) into a form that can be used by the Perceptron Algorithm. The first step of the transformation is to replace the objective function $\max c^T x$ by $c^T x \geq c_0$ for some c_0. We now have:

$$c^T x \geq c_0 \tag{4.7}$$
$$\tilde{a}_i^T x \leq b_i \quad \forall i \tag{4.8}$$

In order to apply the Perceptron Algorithm, we need the right-hand-sides to be 0 and for the inequality to be strict. We therefore replace the linear program (4.7, 4.8) with the constraints:

$$(-c, c_0)^T (y, y_0) < 0 \tag{4.9}$$
$$(\tilde{a}_i, -b_i)^T (y, y_0) < 0 \quad \forall i \tag{4.10}$$

which we then feed into the Perceptron Algorithm (negating both sides to replace "<" with ">").

Constraints (4.9, 4.10) are not identical to (4.7, 4.8) but they are closely related as follows. First, given a solution to the original system (4.7, 4.8) that lies strictly in the interior, we can form a solution to the transformed system (4.9, 4.10) via

$$(y, y_0) = (x, 1)$$

Now suppose we have a solution (y, y_0) to the transformed system (4.9, 4.10) produced by the Perceptron Algorithm. If $y_0 > 0$, then $x = y/y_0$ is a solution to the original system (4.7, 4.8). Similarly, if $y_0 = 0$, then $x = y/\epsilon$ will be a solution to the original for sufficiently small $\epsilon > 0$ (because any solution to (4.9, 4.10) lies in its strict interior). On the other hand, if $y_0 < 0$, and x is any feasible solution in the interior of the feasible region (4.7, 4.8) then $x + \lambda(x - y/y_0)$ is a feasible solution to the linear program (2.5, 2.6) for every $\lambda \geq 0$, and the objective value of this solution increases without bound as we increase λ. Therefore a solution with $y_0 < 0$ provides a certificate that if the linear program (2.5, 2.6) is feasible with objective value at least c_0, it is unbounded.

The event that (4.7, 4.8) is feasible but has empty interior occurs with probability zero. This will follow from lemma 6.2 and letting $\nu \longrightarrow 0$.

We can now assume that the problem we wish to solve is of the form

$$d_j^T w < 0 \quad \forall j \in \{0, \ldots, m\} \tag{4.11}$$
$$w = (y, y_0) \tag{4.12}$$
$$d_0 = (-c, c_0) \tag{4.13}$$
$$d_j = (\tilde{a}_j, -b_j) \quad j \in \{1, \ldots, m\} \tag{4.14}$$

which is a rewriting of the system (4.9, 4.10).

5 Geometric Arguments

We will find the following theorem due originally to Brunn and Minkowski very useful. We prove it in appendix B for completeness.

THEOREM 5.1. (BRUNN-MINKOWSKI) *Let K be a d-dimensional convex body, and let $\bar{x}$ denote the center of mass of K, $\bar{x} = \mathbf{E}_{x\in K}[x]$. Then for every w,*

$$\frac{\max_{x\in K} w^T(x-\bar{x})}{\max_{x\in K} w^T(\bar{x}-x)} \le d$$

To give the reader a feel for the meaning of theorem 5.1, suppose we have a convex body and some hyperplane tangent to it. If the maximum distance from the hyperplane to a point in the convex body is at least t, then the center of mass of the convex body is at least $\frac{t}{d+1}$ away from the bounding hyperplane.

We now state a lemma which will be crucial to our proof of theorem 1.1. We defer the proof to appendix D. No details of the proof of lemma 5.1 are needed for the proof of our main theorem.

LEMMA 5.1. (SMALL BOUNDARIES ARE EASILY MISSED) *Let K be an arbitrary convex body, and let $\Delta(K,\epsilon)$ denote the ϵ-boundary of K, i.e.,*

$$\Delta(K,\epsilon) = \{x : \exists x' \in K, |x-x'| \le \epsilon\} \setminus K$$

Let g be chosen according to a d-dimensional Gaussian distribution with mean $\bar{g}$ and variance σ^2, $g \sim N(\bar{g},\sigma)$. Then

$$\Pr[g \in \Delta(K,\epsilon)] = O\left(\frac{\epsilon\sqrt{d}}{\sigma}\right)$$

6 Proof of the Main Theorem

The next two lemmas will directly imply theorem 1.1. Let $\tilde{M}$ denote the linear programming feasibility problem given by constraints (4.11, 4.12, 4.13, 4.14). $\tilde{M}$ is the linear program $\tilde{L}$ recast as a linear programming feasibility problem in conic form (as explained in section 2).

When $\tilde{M}$ is feasible, we define t_i to be the sine of the maximum angle between any point w' in the feasible region and the hyperplane $(-d_i)^T w \ge 0$, where we view the feasible point w' as a vector from the origin. That is

$$t_i = \max_{w' \text{ feasible for } \tilde{M}} \frac{-d_i^T w'}{|d_i||w'|}$$

This is the same as the cosine between $-d_i$ and w'. Intuitively, if t_i is large, this constraint does not make the feasible region small.

LEMMA 6.1. (MARGIN FOR A SINGLE CONSTRAINT) *Fix $i \in \{1,\ldots,m\}$.*

$$\Pr[\tilde{M} \textit{ is feasible and } t_i \le \epsilon] = O\left(\frac{\epsilon\sqrt{d}}{\sigma}\log\frac{\sigma}{\epsilon\sqrt{d}}\right)$$

Proof. We imagine applying the perturbation to a_i last, after all the $a_j, j \ne i$, have already been perturbed. Let R denote the set of points (in the polar, normal vectors to the hyperplane) w satisfying all the other constraints after perturbation, i.e., $R = \{w : d_j^T w \le 0 \quad \forall j \ne i\}$. No matter what R is, the random choice of perturbation to a_i will be enough to prove the lemma. If R is empty, then we are done, because $\tilde{M}$ will be infeasible no matter what $d_i = (\tilde{a}_i, b_i)$ is. Thus we may assume that R is non-empty.

Define D to be the set of possible values d_i could take on so that $\tilde{M}$ is infeasible, i.e.,

$$D = \{d_i : d_i^T w > 0 \quad \forall w \in R\}$$

Note that D is a convex cone from the origin. We define F to be an "ϵ-boundary" of D in the sense of the sine of the angle between vectors in D and F. That is,

$$F = \{d_i : \exists d_i' \in D \quad s.t. \quad \frac{d_i^T d_i'}{|d_i||d_i'|} \ge \sqrt{1-\epsilon^2}\} \setminus D$$

F is the set of normal vectors d_i to a hyperplane $d_i^T w \le 0$ that could be rotated by an angle whose sine is ϵ or less to some other vector d_i' and yield that $R \cap \{w : {d'}_i^T w \le 0\}$ is empty. F is useful because it is exactly the set of possibilities for d_i that we must avoid if we are to have $t_i > \epsilon$. We justify this claim about F in appendix C.

Because we are not applying a perturbation to the entire vector (a_i, b_i), we are interested in the restriction of D and F to the hyperplane where the $(d+1)^{st}$ coordinate is b_i. Clearly $D \cap \{d_i : d_i[d+1] = b_i\}$ is still convex. However, $F \cap \{d_i : d_i[d+1] = b_i\}$ may contain points that are not within distance $O(\epsilon)$ of $D \cap \{d_i : d_i[d+1] = b_i\}$ (even though $F \cap \{d_i : d_i[d+1] = b_i\}$ is still an "ϵ-boundary" of $D \cap \{d_i : d_i[d+1] = b_i\}$ in the sense of the sine of the angle between two vectors). To overcome this, we condition on the point d_i being a bounded distance away from the origin; then ϵ variation in sine of the angle between two vectors will correspond to a proportional variation in distance. We proceed to make this formal.

We can upper bound the probability that $|\tilde{a}_i - a_i| \ge \kappa$ by analyzing a sum of Gaussians. Since $|(a_i,b_i)| \le \sqrt{2}$ this will give us an easy upper bound of $\kappa + 2$ on $|d_i|$ with the same probability. The following technical statement is proved in appendix A following the outline of Dasgupta and Gupta[8].

FACT 6.1. (SUM OF GAUSSIANS) *Let $X_1,\ldots,X_d$ be independent $N(0,\sigma)$ random variables. Then*

$$\Pr[\sum_{i=1}^{d} X_i^2 \ge \kappa^2] \le e^{\frac{d}{2}(1-\frac{\kappa^2}{d\sigma^2}+\ln\frac{\kappa^2}{d\sigma^2})}$$

Fact 6.1 yields that $\Pr[|d_i| \geq \kappa + 2] \leq e^{-\kappa^2/4}$ for $\kappa \geq 1$ (using that $\sigma^2 \leq 1/2d$). Suppose now that $|d_i| \leq \kappa + 2$. Define

$$D' = D \cap \{d_i : d_i[d+1] = b_i\} \cap \{d_i : |d_i| \leq \kappa + 2\}$$

$$F' = F \cap \{d_i : d_i[d+1] = b_i\} \cap \{d_i : |d_i| \leq \kappa + 2\}$$

Since $|d_i| \leq \kappa+2$, we just need to show $d_i \notin F'$ in order to have $t_i > \epsilon$. Given a point $p_1 \in F'$, $\exists \quad p_2 \in D'$ such that the sine of the angle between the two points is at most ϵ. To show that F' is contained by an $O(\kappa^2\epsilon)$-boundary of D' in the sense of distance, we will show that any two points in $\{d_i : d_i[d+1] = b_i, |d_i| \leq \kappa + 2\}$ at distance $|\gamma|$ from each other satisfy that the sine of the angle between the two points is $\Omega(|\gamma/\kappa^2|)$. To reduce notation (and without loss of generality) assume $b_i = 1$. Let p_1 and p_2 be two points, $p_1 = (p, 1), p_2 = (p + \gamma, 1)$ (where γ is a vector of magnitude $|\gamma|$, and $|p| = O(\kappa)$). Then the sine of the angle we want is given by $\sqrt{1 - \frac{(p_1^T p_2)^2}{p_1^2 p_2^2}}$. We proceed to evaluate

$$\frac{(p_1^T p_2)^2}{p_1^2 p_2^2} = \frac{1 + 2p^2 + 2p^T\gamma + 2p^2 p^T\gamma + p^4 + (p^T\gamma)^2}{1 + 2p^2 + 2p^T\gamma + 2p^2p^T\gamma + p^4 + (p^2\gamma^2) + \gamma^2}$$

$$= \frac{1}{1 + \Omega(\gamma^2/\kappa^4)} = 1 - \Omega(\gamma^2/\kappa^4)$$

Therefore the sine of the angle between p_1 and p_2 is $\Omega(\gamma/\kappa^2)$.

The above discussion has led us to the following simple situation: we are seeking to show that any point subject to a Gaussian perturbation of variance σ^2 has a good chance of missing the $O(\kappa^2\epsilon)$-boundary of a convex body. By lemma 5.1, the perturbed point hits the boundary with probability at most $O(\kappa^2\epsilon\sqrt{d}/\sigma)$. The following calculation illuminates what value to choose for κ to obtain the claimed bound for this lemma. Let H be the event that the perturbed point hits the boundary.

$$\Pr[H] = \Pr[H \mid d_i^2 \leq \kappa^2] \Pr[d_i^2 \leq \kappa^2] + \Pr[H \mid d_i^2 > \kappa^2] \Pr[d_i^2 > \kappa^2]$$

$$\leq O(\kappa^2\epsilon\sqrt{d}/\sigma) \cdot 1 + 1 \cdot e^{-\kappa^2/4}$$

Setting $\kappa^2 = \log(\sigma/(\epsilon\sqrt{d}))$ concludes the proof of lemma 6.1. ∎

We now turn to the last lemma we want for the proof of our main theorem. The idea of the lemma is that if no single constraint leads to a small margin (small t_i), then the Brunn-Minkowski theorem will imply that the feasible region contains a solution with large wiggle room. A simple trick allows us to get away with perturbing all but one of the constraints (rather than all).

LEMMA 6.2. (MARGIN FOR MANY CONSTRAINTS) *Let E denote the event that $\tilde{M}$ is feasible yet contains no solution of wiggle room ν.*

$$\Pr[E] = O(\frac{md^{1.5}\nu}{\sigma} \log \frac{\sigma}{d^{1.5}\nu})$$

Proof. Setting $\epsilon = 4(d+1)\nu$, it is a straightforward application of the union bound and lemma 6.1 that

$$\Pr[\tilde{M} \text{ is feasible and yet for some } i,\, t_i \leq \epsilon]$$

$$= O(\frac{md^{1.5}\nu}{\sigma} \log \frac{\sigma}{d^{1.5}\nu})$$

We now show that if for every i, $t_i > \epsilon$, then the feasible region $\tilde{M}$ contains a vector w' with wiggle room ν. If the reader desires to visualize w' with wiggle room ν, we suggest picturing that w' forms the axis of an ice cream cone lying entirely in $\tilde{M}$, where any vector along the boundary of the ice cream cone is at an angle from w' whose sine is ν. Because the Brunn-Minkoski theorem applies to distances, not angles, we will consider the restriction of our feasible cone to a hyperplane.

Let w^* be the unit vector that satisfies $\tilde{M}$ with maximum wiggle room, and denote the wiggle room by ν'. We suppose for purpose of contradiction that $\nu' < \nu$. Consider the restriction of the $(d+1)$-dimensional cone $\tilde{M}$ to the d-dimensional hyperplane $\tilde{M}'$ defined by

$$\tilde{M}' = \tilde{M} \cap \{w : w^T w^* = 1\}$$

$\tilde{M}'$ is clearly convex. In $\tilde{M}'$, w^* forms the center of a sphere of radius $R = \frac{\nu'}{\sqrt{1-\nu'^2}} \leq 2\nu'$ for $\nu' \leq 1/2$ (if $\nu' > 1/2$, we are done). The restriction to $\tilde{M}'$ maintains the contact between the boundary of the ice cream cone and the bounding constraints, so w^* forms the center of a sphere of maximum radius over all spheres lying within $\tilde{M}'$.

Let H_i be the hyperplane $d_i^T w = 0$, and let H_i' be H_i restricted to $\{w : w^T w^* = 1\}$. Define $s_i = \max\{\text{distance of } w' \text{ to } H' : w' \in \tilde{M}'\}$. We now show $s_i \geq t_i, \forall i$. Fix i, and let $\hat{w} \in \tilde{M}$ be a unit vector satisfying $\frac{-d_i^T \hat{w}}{|d_i|} = t_i$. Then $\hat{w}$ is exactly distance t_i from the hyperplane H_i. Let $\hat{w'}$ be a scalar multiple of $\hat{w}$ such that $\hat{w'} \in \tilde{M}'$. The norm of $\hat{w'}$ is at least that of $\hat{w}$, and so $\hat{w'}$ is distance at least t_i from H_i. Since $\hat{w'}$ is distance at least t_i from H_i, it is distance at least t_i from H_i' (using that H_i' is a restriction of H_i). Thus $s_i \geq t_i$.

Let $\bar{w} = \mathbf{E}_{w\in\tilde{M}'}[w]$, the center of mass of $\tilde{M}'$. We apply theorem 5.1 to conclude that $\bar{w}$ is distance at least $\frac{s_i}{d+1} \geq 4\nu$ from the i^{th} constraint, H_i', for all $i \in \{1, \dots m\}$. We now consider the unperturbed constraint, $d_0^T w \leq 0$. Since $\bar{w}$ satisfies $d_0^T w \leq 0$, we construct $\bar{w}'$ by starting at $\bar{w}$, and then moving a distance 2ν away from the restriction of the hyperplane $d_0^T w = 0$ to $\{w : w^T w^* = 1\}$. Since $\bar{w}$ was distance at least 4ν from all the other hyperplanes $H_i', i \in \{1, \dots, m\}$, $\bar{w}'$ is distance at least 2ν from all the other hyperplanes H_i'. An explicit formula for $\bar{w}'$ is given by $\bar{w}' = \bar{w} - 2\nu d_0'/|d_0'|$, where $d_0' = d_0 - (d_0^T \bar{w})\bar{w}$. We conclude that $\bar{w}'$ is the center of a radius 2ν sphere lying entirely within $\tilde{M}'$, contradicting the assumption that the sphere of maximum radius in $\tilde{M}'$ had radius at most $2\nu' < 2\nu$. This concludes the proof of lemma 6.2.

■

Proof. (of Theorem 1.1) Lemma 6.2 and theorem 3.1 are enough to conclude that for fixed c_0, we can identify a solution x satisfying $c^T x \geq c_0$ as in theorem 1.1. Set $\nu = O(\frac{\delta\sigma}{md^{1.5}\ln(m/\delta)})$ and then with probability at least $1 - \delta$, either we find a solution to $\tilde{M}$ in $O(1/\nu^2)$ iterations, or $\tilde{M}$ is infeasible. If $\tilde{M}$ is infeasible, then $\tilde{L}$ is infeasible. If we find a solution (y, y_0) to $\tilde{M}$ with $y_0 > 0$, we have a solution to $\tilde{L}$ with objective value at least c_0. If $y_0 < 0$, we know that $\tilde{L}$ is either infeasible for the chosen value of c_0 or unbounded. ■

7 Discussion

The first observation we make is that the preceding analysis was tailored to show that the perceptron algorithm works in the exact same model of perturbation that Spielman and Teng used. Our analysis would have been shorter if our model of perturbation had instead been the following: Start with a system of linear inequalities $\{d_j^T w \leq 0\}$ for which we want to find a feasible point. Then perturb each d_j by rotating it a small random amount in a random direction.

The second observation we make concerns the issue of what polynomial running time in the smoothed complexity model suggests about the possibility of strongly polynomial running time in the standard complexity model. The ellipsoid algorithm and interior-point methods are not strongly polynomial, while one of the appealing aspects of the simplex algorithm is the possibility that a strongly polynomial pivot rule will be discovered. The analysis in this paper suggests that the smoothed complexity model sweeps issues of bit size under the rug, as the following analysis of the ellipsoid algorithm makes clear.

In the ellipsoid algorithm, we start with a ball of radius 2^L, where L is a function of the encoding length of the input, and is polynomially related to the bit size. Separating hyperplanes are then found until the algorithm has obtained a feasible point, or else ruled out every region of radius greater than 2^{-L}. In the proof of theorem 1.1, we transformed the linear program so that the desired solution was now a vector w in $d+1$ dimensional space, and every scalar multiple of w was equivalent. Consider a regular simplex around the origin, scaled so that it contains a unit ball, and let each of the $d+2$ faces represent a different possible plane to which we could restrict the ellipsoid algorithm. Each face is contained by a d dimensional ball of radius $d+2$. If the problem is feasible, one of the $d+2$ faces contains a ball of radius $\tilde{O}(\frac{\sigma\delta}{md^{1.5}})$ with good probability. Therefore the ellipsoid algorithm runs in expected time polynomial in m, d, and $\log 1/\sigma$, with no reference at all to L.

Our main theorem suggests that we should commonly observe the perceptron algorithm to outperform the simplex algorithm, yet in practice, the simplex algorithm is much more widely used than the perceptron algorithm for the task of solving linear programs. (The use of the perceptron algorithm in machine learning is due in large part to other needs in that area, such as behaving reasonably even when the linear program is infeasible.) We offer several possible explanations for the disparity between our theoretical analysis and the observed performance in practice. One possibility is that the simplex algorithm has much better smoothed complexity than is naively inferable from the bound cited at the beginning of this paper. Another possibility is that the perceptron algorithm's failure to achieve the optimum of a particular perturbed linear program is a noticeable hindrance in practice. Yet a third possibility is that a different model of perturbation is needed to distinguish between the observed performance of the simplex and perceptron algorithms. If this last statement were the case, a *relative perturbation* model, such as that put forward by Spielman and Teng in [1], seems to offer a promising framework. It seems that the polynomial time guarantee for the perceptron algorithm would not stand up to this *relative smoothed analysis*, while the simplex algorithm well might still have polynomial running time.

References

[1] D. Spielman, S. Teng, "Smoothed Analysis: Why The Simplex Algorithm Usually Takes Polynomial Time," In *Proc. of the 33rd ACM Symposium on the Theory of Computing*, Crete, 2001.

[2] F. Rosenblatt. *Principles of Neurodynamics*, Spartan Books, 1962.

[3] S. Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6(3):382-392, 1954.

[4] T. Bylander. Polynomial learnibility of linear threshold approximations. In *Proceedings of the Sixth Annual Workshop on Computational Learning Theory*, pages 297-302. ACM Press, New York, NY, 1993.

[5] T. Bylander. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the Seventh Annual Workshop on Computational Learning Theory*, pages 340-347. ACM Press, New York, NY, 1994.

[6] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, pages 392-401, 1993.

[7] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry.* The MIT Press, 1969.

[8] S. Dasgupta, A. Gupta. An elementary proof of the Johnson-Lindenstrauss Lemma. International Computer Science Institute, Technical Report 99-006.

[9] R. J. Gardner. The Brunn-Minkowski Inequality. `http://www.ac.wwu.edu/~gardner/` Submitted for publication.

[10] P. M. Gruber, J. M. Wills, editors. *Handbook of convex geometry*, chapter 1.2. Elsevier Science Publishers, 1993.

[11] K. Ball, "The Reverse Isoperimetric Problem for Gaussian Measure," in *Discrete and Computational Geometry*, vol. 10 no. 4, pp. 411-420, 1993.

[12] Bhattacharya and Rao, Normal Approximation and Asymptotic Expansion, pp. 23-38, 1976.

A Bounds on Sum of Gaussians

We restate the bound on a sum of Gaussians (fact 6.1) that we previously deferred proving. The distribution we are analyzing is the Chi-Squared distribution, and bounds of this form are well-known.

FACT A.1. (SUM OF GAUSSIANS) *Let $X_1, \dots, X_d$ 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

$$\begin{aligned}
\Pr[\sum_{i=1}^{d} Y_i^2 \geq k] &= \\
\Pr[\sum_{i=1}^{d} Y_i^2 - k \geq 0] &= && \text{(for } t > 0) \\
\Pr[e^{t(\sum_{i=1}^{d} Y_i^2 - k)} \geq 1] &\leq && \text{(by Markov's Ineq.)} \\
\mathbf{E}[e^{t(\sum_{i=1}^{d} Y_i^2 - k)}] &= \\
\left(\frac{1}{1-2t}\right)^{d/2} e^{-kt} &\leq && \text{(letting } t = \frac{1}{2} - \frac{d}{2k}) \\
\left(\frac{k}{d}\right)^{d/2} e^{-\frac{k}{2}+\frac{d}{2}} &= e^{\frac{d}{2}(1-\frac{k}{d}+\ln\frac{k}{d})}
\end{aligned}$$

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 A.2. (ALTERNATIVE SUM OF GAUSSIANS) *Let $X_1, \dots, X_d$ be independent $N(0, \sigma)$ random variables. Then*

$$\Pr[\sum_{i=1}^{d} X_i^2 \geq cd\sigma^2] \leq e^{\frac{d}{2}(1-c+\ln c)}$$

$$\Pr[\sum_{i=1}^{d} X_i^2 \leq cd\sigma^2] \leq e^{\frac{d}{2}(1-c+\ln c)}$$

Proof. The first inequality is proved by setting $k = cd$ in the last line of the proof of fact A.1. To prove the second inequality, begin the proof of fact A.1 with $\Pr[\sum_{i=1}^{d} Y_i^2 \leq k]$ and continue in the obvious manner. ∎

B Proof of Brunn-Minkowski Theorem

We restate theorem 5.1 and then prove it. This theorem is one of many results belonging to the Brunn-Minkowski theory of convex bodies.

THEOREM B.1. (BRUNN-MINKOWSKI) *Let K be a d-dimensional convex body, and let $\bar{x}$ denote the center of mass of K, $\bar{x} = \mathbf{E}_{x \in K}[x]$. Then for every w,*

$$\frac{\max_{x \in K} w^T(x - \bar{x})}{\max_{x \in K} w^T(\bar{x} - x)} \leq d$$

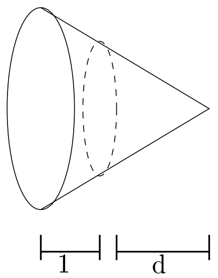

Figure 1: Worst case K for theorem 5.1.

Proof. The entire proof consists of showing that figure 1 is the worst case for the bound we want. Without loss of generality, let $\bar{x}$ be the origin. Let K and w be fixed, and let w be a unit vector. Consider the body K' that is rotationally symmetric about w and has the same $(d-1)$-dimensional volume for every cross section $K_r = \{x : x \in K, w^T x = r\}$, i.e., $vol_{d-1}(K_r) = vol_{d-1}(K'_r)$. K' is referred to as the *Schwarz rounding* of K in [10]. K' has the same mean as K, and also the same min and max as K when we consider the projection along w, but K' will be easier to analyze. Denote the radius of the $(d-1)$-dimensional ball K'_r by $radius(K'_r)$. That K' is convex follows from the Brunn-Minkowski inequality

$$vol_n((1-\lambda)A+\lambda B)^{1/n} \geq (1-\lambda)vol_n(A)^{1/n}+(\lambda)vol_n(B)^{1/n}$$

where A and B are convex bodies in $\mathcal{R}^n$, $0 < \lambda < 1$, and $+$ denotes the Minkoski sum. Proofs of this inequality can be found in both [9] and [10]. To see the implication of the theorem from the inequality, let A, B be two cross sections of K, $A = K_{r_1}$, $B = K_{r_2}$ and consider the cross-section $K_{(r_1+r_2)/2}$. By convexity of K, $\frac{1}{2}A + \frac{1}{2}B \subset K_{(r_1+r_2)/2}$, and therefore

$$vol_{d-1}(K_{(r_1+r_2)/2})^{1/(d-1)} \geq \frac{1}{2}vol_{d-1}(K_{r_1})^{1/(d-1)} + \frac{1}{2}vol_{d-1}(K_{r_2})^{1/(d-1)}$$

This implies that $radius(K'_{(r_1+r_2)/2}) \geq \frac{1}{2}radius(K'_{r_1}) + \frac{1}{2}radius(K'_{r_2})$, which yields that K' is convex.

Let $radius(K'_0) = R$, and let $[\max w^T(x - \bar{x})] = r_0$. Then $radius(K'_r) \geq R(1 - \frac{r}{r_0})$ for $r \in [0, r_0]$ by convexity. Similarly, $radius(K'_r) \leq R(1 - \frac{r}{r_0})$ 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 $r_1 = [\max w^T(\bar{x} - x)]$ 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 $r_1 = \frac{r_0}{d}$. ∎

C Justification for Definition of F

We justify here that for F and D defined as in section 6, F is exactly the set of vectors such that $t_i \leq \epsilon$.

Let d_i be a fixed unit vector. We first show that $t_i > \epsilon \Rightarrow d_i \notin F$. Let $w' \in \tilde{M}$ be a point realizing the maximum t_i. Every $d'_i \in D$ must make w' infeasible, and so every $d'_i \in D$ is more than ϵ away from d_i (by more than ϵ away, we mean that the sine of the angle between d'_i and d_i is at least ϵ). Thus $d_i \notin F$. Now we show that $t_i \leq \epsilon \Rightarrow d_i \in F$. The proof uses that $\tilde{M}$ is a convex cone. Let $w' \in \tilde{M}$ be a point that realizes the maximum $t_i, t_i \leq \epsilon$. We claim that rotating the hyperplane d_i in the direction of w' by the amount t_i will make $\tilde{M}$ empty (and thus d_i is within ϵ of $d'_i \in D$). Another way to say this is that $d'_i = d_i/|d_i| + t_i w'/|w'|$ is in D. It is clear that d'_i is within t_i of d_i (i.e., the sine of the angle between d_i and d'_i is t_i). To verify that $d'_i \in D$, suppose it were not true, i.e., there were some point $\tilde{w} \in \tilde{M}$ that is feasible for the rotated hyperplane d'_i. Then we show that $\tilde{w}$ and w' define a cone containing some point (also in $\tilde{M}$) more than t_i away from the unrotated d_i (i.e., the sine of the angle between the constructed point and d_i is more than t_i). This will contradict our assumption about t_i equaling $\max\{\frac{-d_i^T w'}{|d_i||w'|} : w' \text{ feasible for } \tilde{M}\}$. Let $-{d'}_i^T \tilde{w} = c > 0$ (since $\tilde{w}$ is feasible for d'_i), and construct $p = \alpha\tilde{w} + w'$. Because $\tilde{M}$ is a convex cone, $p \in \tilde{M}$. We seek to find $\alpha > 0$ such that $\frac{-d_i^T p}{|d_i||p|} > t_i$. We expand the left hand side of the desired inequality as

$$\begin{aligned}
\frac{-d_i^T p}{|d_i||p|} &= \frac{-(d'_i - t_i w'/|w'|)^T(\alpha\tilde{w} + w')}{\sqrt{\alpha^2\tilde{w}^2 + 2\alpha\tilde{w}^T w' + |w'|^2}} \\
&= \frac{\alpha c + \alpha t_i \tilde{w}^T w'/|w'| + t_i|w'|}{\sqrt{\alpha^2\tilde{w}^2 + 2\alpha\tilde{w}^T w' + |w'|^2}} \\
&\geq \frac{\alpha c + \alpha t_i \tilde{w}^T w'/|w'| + t_i|w'|}{\alpha^2\tilde{w}^2/(2|w'|) + \alpha\tilde{w}^T w'/|w'| + |w'|}
\end{aligned}$$

We see that as α approaches 0, but before α reaches 0, the quantity on the right-hand side of the above expression is strictly greater than t_i. This completes the argument that $t_i \leq \epsilon \Rightarrow d_i \in F$.

D Proof that Small Boundaries are Easily Missed

Before proving lemma 5.1, we prove fact D.1, which will be useful in proving lemma 5.1.

FACT D.1. (SURFACE AREA OF A CONVEX BODY) *Let A be a convex body in $\mathcal{R}^d$, $A \subset B$. Denote the boundary of a region R by $\Delta(R)$. Then*

$$vol_{d-1}(\Delta(A)) \leq vol_{d-1}(\Delta(B))$$

Proof. Because A is convex, we can imagine transforming B into 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 A. ∎

We restate lemma 5.1 and then prove it.

LEMMA D.1. (SMALL BOUNDARIES ARE EASILY MISSED) *Let K be an arbitrary convex body, and let $\Delta(K,\epsilon)$ denote the ϵ-boundary of K, i.e.,*

$$\Delta(K,\epsilon) = \{x : \exists x' \in K, |x - x'| \le \epsilon\} \setminus K$$

Let g be chosen according to a d-dimensional Gaussian distribution with mean $\bar{g}$ and variance σ^2, $g \sim N(\bar{g},\sigma)$. Then

$$\Pr[g \in \Delta(K,\epsilon)] = O\left(\frac{\epsilon\sqrt{d}}{\sigma}\right)$$

Proof.

This bound is tight to within a factor of $\Theta(\sqrt{d})$, as can be seen from letting K be a hyperplane passing through $\bar{g}$. For the proof, we divide space into thin shells of a hypersphere (like an onion) centered at $\bar{g}$. 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, $\Delta(K,\epsilon)$ can't be more than a small fraction of the overall volume of the shell.

Without loss of generality, let $\bar{g}$ be the origin. Recall that the probability density function of g is given by

$$\mu(x) = \left(1/\sqrt{2\pi}\right)^d e^{-|x|^2/2}$$

As before, let $\Delta(X)$ denote the boundary of the region X. Fix $\gamma > 0$.
Let $S_R = \{x : R \le |x| \le (1 + \frac{\gamma}{d})R\}$.

We would like to be able to argue that, if $\Delta(K,\epsilon)$ is a small fraction of the volume of S_R, then if we condition on g landing within S_R, we are unlikely to land in $\Delta(K,\epsilon)$. The concept of *bias* allows us to make this argument. Define the *bias* of a region X by

$$bias(X) = \frac{\max_{x\in X}\mu(x)}{\min_{x\in X}\mu(x)}$$

Then we can say that, for any $Y \subset X$,

$$\Pr[g \in Y | g \in X] \le \frac{vol(Y)}{vol(X)} \cdot bias(X)$$

For S_R, we calculate

$$bias(S_R) = \frac{e^{-R^2/\sigma^2}}{e^{-(1+\gamma/d)^2R^2/\sigma^2}} = e^{(2\gamma/d+\gamma^2/d^2)R^2/\sigma^2}$$

We upper bound the probability of landing in $\Delta(K,\epsilon)$ using

$$\Pr[g \in \Delta(K,\epsilon) | g \in S_R] \le \frac{vol(\Delta(K,\epsilon)\cap S_R)}{vol(S_R)} \cdot bias(S_R)$$

Let B be a ball of radius $(1+\frac{\gamma}{d})R$. Let K' be the convex closure of $\Delta(K,\epsilon)\cap S_R$. Clearly $K' \subset B$. We can upper bound $vol(\Delta(K,\epsilon)\cap S_R)$ by $\epsilon\cdot vol_{d-1}(\Delta(K'))$, and by fact D.1, this is at most $\epsilon\cdot vol_{d-1}(B)$. The exact formulas for the volume and surface area of a sphere are

$$vol(S_R) = \frac{2((1+\frac{\gamma}{d})R)^d\pi^{d/2}}{d\Gamma(d/2)} - \frac{2R^d\pi^{d/2}}{d\Gamma(d/2)}$$

$$vol_{d-1}(B) = \frac{2((1+\frac{\gamma}{d})R)^{d-1}\pi^{d/2}}{\Gamma(d/2)}$$

which yields

$$\frac{vol(\Delta(K,\epsilon)\cap S_R)}{vol(S_R)} bias(S_R) \le$$

$$\frac{d\epsilon}{R}\cdot\frac{(1+\frac{\gamma}{d})^{d-1}}{(1+\frac{\gamma}{d})^d - 1}\cdot e^{\frac{\gamma}{d}(1+\gamma/d)^2(2+\gamma/d)R^2/\sigma^2}$$

To complete the proof, we sum over all the possible shells S_R that g might land in. This is done in the following formula.

$$\Pr[g \in \Delta(K,\epsilon)] \le \sum_{k,R=(1+\frac{\gamma}{d})^k} \Pr[g \in \Delta(K,\epsilon) \mid g \in S_R]\Pr[g \in S_R]$$

$$\le \sum_{k,R=(1+\frac{\gamma}{d})^k} \Pr[g \in S_R]\cdot\frac{d\epsilon}{R}\cdot\frac{(1+\frac{\gamma}{d})^{d-1}}{(1+\frac{\gamma}{d})^d - 1}\cdot e^{\frac{\gamma}{d}(1+\gamma/d)^2(2+\gamma/d)R^2/\sigma^2}$$

$$\le \mathbf{E}_{\{g,|g|=\sigma\sqrt{cd}\}}\left[\frac{\sqrt{d}\epsilon}{\sqrt{c}\sigma}\cdot\frac{(1+\frac{\gamma}{d})^d}{(1+\frac{\gamma}{d})^d - 1}\cdot e^{\gamma(1+\gamma/d)^4(2+\gamma/d)c}\right]$$

We use the identity
$\mathbf{E}_g[f(g)] = \int_{x=0}^{\infty}\Pr_g[f(g) > x]dx$ 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:

$$\mathbf{E}[\frac{1}{\sqrt{c}}e^{2\gamma_2 c}] = \int_{x=0}^{\infty} \Pr_{\{g,|g|=\sigma\sqrt{cd}\}}[\frac{1}{\sqrt{c}}e^{2\gamma_2 c} > x]dx$$

$$= \int_{x=0}^{\infty} \Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2 c} > x,\ c \geq 1] + \Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2 c} > x,\ c < 1]dx$$

$$\leq \int_x \Pr[e^{2\gamma_2 c} > x \text{ and } c \geq 1] + \Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2} > x \text{ and } c < 1]dx$$

$$= \int_{x=e^{2\gamma_2}}^{\infty} \Pr[e^{2\gamma_2 c} > x]dx + \int_{x=e^{2\gamma_2}}^{\infty} \Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2} > x]dx$$

$$= \int_{x=e^{2\gamma_2}}^{\infty} \Pr[c > \frac{\ln x}{2\gamma_2}]dx + \int_{x=e^{2\gamma_2}}^{\infty} \Pr[c < \frac{e^{4\gamma_2}}{x^2}]dx$$

$$\leq \int_{x=e^{2\gamma_2}}^{\infty} e^{\frac{d}{2}(1-c'+\ln c')}|_{c'=\frac{\ln x}{2\gamma_2}}dx + \int_{x=e^{2\gamma_2}}^{\infty} e^{\frac{d}{2}(1-c'+\ln c')}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx$$

$$\leq \int_{x=e^{2\gamma_2}}^{\infty} e^{(1-c'+\ln c')}|_{c'=\frac{\ln x}{2\gamma_2}}dx + \int_{x=e^{2\gamma_2}}^{\infty} e^{(1-c'+\ln c')}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx$$

Where on the last step we observe that $1 - c' + \ln c' \leq 0$ and we assume that $d \geq 2$. We now proceed to analyze the right-hand term.

$$\begin{aligned}\int_{x=e^{2\gamma_2}}^{\infty} e^{(1-c'+\ln c')}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx &\leq \int_{x=e^{2\gamma_2}}^{\infty} e^{1+\ln c'}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx \\ &= e\int_{x=e^{2\gamma_2}}^{\infty} \frac{e^{4\gamma_2}}{x^2}dx \\ &= e^{2\gamma_2+1}\end{aligned}$$

For the lefthand term we make the change of variables $x = e^{2\gamma_2\alpha}$. Continuing:

$$\begin{aligned}\int_{x=e^{2\gamma_2}}^{\infty} e^{(1-c'+\ln c')}|_{c'=\frac{\ln x}{2\gamma_2}}dx &= \int_{\alpha=1}^{\infty} e^{1-\alpha+\ln\alpha}2\gamma_2 e^{2\gamma_2\alpha}d\alpha \\ &= 2\gamma_2 e\int_{\alpha=1}^{\infty} \alpha e^{(2\gamma_2-1)\alpha}d\alpha \\ &= 2\gamma_2 e\left[\frac{\alpha}{2\gamma_2-1}e^{(2\gamma_2-1)\alpha} - \frac{1}{(2\gamma_2-1)^2}e^{(2\gamma_2-1)\alpha}\right]_{\alpha=1}^{\infty} \\ \text{(since } \gamma_2 < 1/2) &= 2\gamma_2 e^{2\gamma_2}\left[\frac{1}{(2\gamma_2-1)^2} - \frac{1}{2\gamma_2-1}\right]\end{aligned}$$

Our final bound on $\Pr[g \in \Delta(K,\epsilon)]$ is thus

$$\frac{\sqrt{d}\epsilon}{\sigma}\frac{e^{2\gamma_2}}{\gamma_1}\left(e + \frac{4(\gamma_2-\gamma_2^2)}{(2\gamma_2-1)^2}\right)$$

Letting $\gamma = .1$, we derive that this is at most $45\frac{\sqrt{d}\epsilon}{\sigma}$. As d increases, the constant quickly drops off. This concludes the lemma proof.

∎

We thank Ryan O'Donnell for directing us to two previously published proofs of this fact in the literature, [11], [12]. In those proofs, the constant 45 is replaced by 1. Additionally, [11] proves the stronger statement that

Theorem D.1. (K. Ball)

$$\Pr[g \in \Delta(K,\epsilon)] \leq 4\left(\frac{\epsilon d^{1/4}}{\sigma}\right)$$

It is straightforward to use this stronger bound to obtain our main theorem with $\tilde{O}(\frac{m^2 d^{2.5}}{\sigma^2\delta^2})$ in place of $\tilde{O}(\frac{m^2 d^3}{\sigma^2\delta^2})$. Additionally, Ryan O'Donnell communicated to us that F. Nazarov has proved a matching lower bound for theorem D.1.