A Combinatorial Perspective on the Optimization of Shallow ReLU Networks

Michael Matena
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
mmatena@cs.unc.edu

Colin Raffel
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
craffel@cs.unc.edu

Abstract

The NP-hard problem of optimizing a shallow ReLU network can be characterized as a combinatorial search over each training example's activation pattern followed by a constrained convex problem given a fixed set of activation patterns. We explore the implications of this combinatorial aspect of ReLU optimization in this work. We show that it can be naturally modeled via a geometric and combinatoric object known as a zonotope with its vertex set isomorphic to the set of feasible activation patterns. This assists in analysis and provides a foundation for further research. We demonstrate its usefulness when we explore the sensitivity of the optimal loss to perturbations of the training data. Later we discuss methods of zonotope vertex selection and its relevance to optimization. Overparameterization assists in training by making a randomly chosen vertex more likely to contain a good solution. We then introduce a novel polynomial-time vertex selection procedure that provably picks a vertex containing the global optimum using only double the minimum number of parameters required to fit the data. We further introduce a local greedy search heuristic over zonotope vertices and demonstrate that it outperforms gradient descent on underparameterized problems.

1 Introduction

Neural networks have become commonplace in a variety of applications. They are typically trained to minimize a loss on a given dataset of labeled examples using a variant of stochastic gradient descent. However, our theoretical knowledge of neural networks and their training lags behind their practical developments.

Single-layer ReLU networks are an appealing subject for theoretical study. The universal approximation theorem guarantees their expressive power while their relative simplicity makes analysis tractable (Hornik, 1991). We restrict ourselves in this paper to studying empirical risk minimization (ERM) as was done in previous works (Du et al., 2018; Oymak & Soltanolkotabi, 2020), which is justified since the train set performance tends to upper bound the test set performance. Furthermore, modern neural networks achieve zero training loss but nevertheless generalize well (Kaplan et al., 2020; Nakkiran et al., 2021). Minimizing the training loss of a shallow ReLU network is a nonconvex optimization problem. Finding its global minima is difficult and can in fact be shown to be NP-hard in general (Goel et al., 2020). Arora et al. (2016) provide an explicit algorithm for finding the global minima by solving a set of convex optimization problems; however, the size of this set is exponential in both the input dimension d and the number of hidden units m.

In this paper, we explore the combinatorial structure implicit in the global optimization algorithm of Arora et al. (2016). We start by using tools from polyhedral geometry to characterize the set of convex optimization problems and describe the relationships between the subproblems. Notably, we are

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

able to create a special type of polytope called a zonotope (McMullen, 1971) whose vertices have a one-to-one correspondence with the convex subproblems and whose faces represent information about their relationships. We then explore the combinatorial optimization problem implicit in shallow ReLU network empirical risk minimization using the zonotope formalism to help interpret our findings and assist in some proofs.

Since the computational complexity of optimization problems shapes our approach to solving them, we examine the reductions of NP-hard problems introduced in Goel et al. (2020). The datasets produced have examples that are not in general position (i.e. they have nontrivial affine dependencies), which differs from most real-world datasets. We prove that the global optimum of the loss of a shallow ReLU network over such a dataset can have a discontinuous jump for arbitrarily small perturbations of the data, which has a very natural interpretation in our zonotope formalism. This means that a proof of the NP-hardness of ReLU optimization given training examples in general position does not follow from the results Goel et al. (2020) via a simple continuity argument. We therefore present a modification of their proof that uses a dataset in general position.

In contrast to the NP-hardness of general ReLU optimization, sufficient overparameterization allows gradient descent to provably converge to a global optimum in polynomial time, as demonstrated by (Du et al., 2018; Oymak & Soltanolkotabi, 2020). We interpret the proof methods generally used in these works as asserting that sufficient overparameterization allows gradient descent to bypass much of the combinatorial search over zonotope vertices by having a randomly chosen vertex be close to one with a good solution with high probability. We then introduce a novel algorithm that finds a good zonotope vertex in polynomial time requiring only about twice the minimum number of hidden units required to fit the dataset.

Finally, we explore how gradient descent interacts with this combinatorial structure. We provide empirical evidence that it can perform some aspects of combinatorial search but present an informal argument that it is suboptimal. We reinforce this claim by showing that a greedy local search heuristic over the vertices of the zonotope outperforms gradient descent on some toy synthetic problems and simplifications of real-world tasks. In contrast to the NP-hard worst case, these results suggest that the combinatorial searches encountered in practice might be relatively tractable.

We summarize our contributions as follows.

- We are the first to provide an in-depth exposition of the combinatorial structure arising from the set of feasible activation patterns that is implicit in shallow ReLU network optimization. In particular, we show that this structure can be characterized exactly as a Cartesian power of the zonotope generated by the set of training examples.
- We use this formalism to prove necessary conditions for the global optimum of a shallow ReLU network to be discontinuous with respect to the training dataset. We show that this implies that previous NP-hardness proofs of ReLU optimization do not automatically apply to datasets satisfying realistic assumptions, which we rectify by presenting a modification that uses a dataset in general position.
- We explore the role that combinatorial considerations play in the relationship between overparameterization and optimization difficulty. In particular, we introduce a novel polynomial-time algorithm fitting a generic dataset using twice the minimum number of parameters needed.
- We introduce a novel heuristic algorithm that performs a greedy search along edges of a zonotope and show that it outperforms gradient descent on some toy datasets.

We hope that the tools we introduce are generally useful in furthering our understanding of ReLU networks. Notably, they have deep connections to several well-established areas of mathematics (McMullen, 1971; Richter-Gebert & Ziegler, 2017; Ziegler, 2012), which might allow researchers to quickly make new insights by drawing upon existing results in those fields.

2 Empirical Risk Minimization for ReLU Networks

A single ReLU layer consists of an affine transformation followed by a coordinate-wise application of the ReLU nonlinearity $\phi(x) = \max\{x, 0\}$. We can represent an affine transformation from $\mathbb{R}^d \to \mathbb{R}^m$ by an $m \times (d+1)$ matrix by representing its inputs in homogeneous coordinates, i.e. by appending a $(d+1)$-th coordinate to network inputs that is always equal to 1. Hence a single ReLU layer with parameters W can be written as $f_W(\mathbf{x}) = \phi(W\bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ denotes $\mathbf{x}$ expressed

in homogeneous coordinates. A single hidden layer ReLU network consists of a single ReLU layer followed by an affine transformation. We focus on the case of a network with scalar output, so the second layer can be represented by a vector $\mathbf{v} \in \mathbb{R}^{m+1}$. Although the second layer parameters are trained jointly with the first layer in practice, we assume that they are fixed for our analysis. This parallels simplifying assumptions made in previous work (Du et al., 2018).

Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a convex loss function such as MSE or cross-entropy. Since the second layer parameters are fixed, we can incorporate them into a modified loss function $\tilde{\ell} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ given by $\tilde{\ell}(\mathbf{z}, y) = \ell(\mathbf{v}^T \bar{\mathbf{z}}, y)$ that operates directly on the first layer's activations $\mathbf{z}$. This modified loss function is convex since it is the composition of a convex function with an affine function.

Suppose we are given $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ as the training dataset with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for all $i = 1, \ldots, N$. Throughout this paper, we assume that $N > d + 1$. Sometimes we will represent a dataset by a matrix $X \in \mathbb{R}^{(d+1)\times N}$ with each column corresponding to an example and its labels as the vector $\mathbf{y} \in \mathbb{R}^N$. We say that $\mathcal{D}$ is in general position if there exist no nontrivial affine dependencies between the columns of X. The empirical loss $L(W)$, also known as the empirical risk, is defined as the mean per-example loss

$$L(W) = \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}(f_W(\mathbf{x}_i), y_i). \tag{1}$$

The goal of ERM is to find a set of parameters W that minimizes this loss. Arora et al. (2016) were the first to introduce an algorithm for exact ERM. We adapt their algorithm for the case of fixed second layer weights in algorithm 1, which has a running time of $O(N^{md} \operatorname{poly}(N, m, d))$. The algorithm works by iterating over all feasible activation patterns

$$\mathcal{A} = \left\{ \mathbb{I}\{W\bar{X} > 0\} \in \{0,1\}^{m\times N} \mid W \in \mathbb{R}^{m\times(d+1)} \right\}. \tag{2}$$

The subset of parameters corresponding to an activation pattern, which we call an activation region, can be expressed via a set of linear inequalities. Within a single activation region, the map from parameter values to ReLU layer activations over the training dataset is linear. Hence we can solve a constrained convex optimization problem to get the optimal parameters in each activation region. Namely for a given $A \in \mathcal{A}$, we solve for $W \in \mathbb{R}^{m\times(d+1)}$ in the following

$$\begin{aligned} \text{minimize} \quad & \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}(\mathbf{a}_i \odot (W\bar{\mathbf{x}}_i), y_i) \\ \text{subject to} \quad & (2\mathbf{a}_i - 1) \odot (W\bar{\mathbf{x}}_i) \geq 0 \end{aligned} \tag{3}$$

where $\odot$ denotes the Hadamard product and $\mathbf{a}_i$ denotes the i-th column of A. The global optimum then becomes the best optimum found over the entire set of activation regions. Thus single-layer ReLU network ERM can be expressed as a combinatorial search over activation patterns with a convex optimization step per pattern.

3 Zonotope Formalism

While Arora et al. (2016) mention that $\mathcal{A}$ arises from a set of hyperplanes induced by the training examples, they only use this connection to bound its cardinality $|\mathcal{A}| = O(N^{md})$. However, hyperplane arrangements are well-studied geometric and combinatoric objects (Richter-Gebert & Ziegler, 2017; Stanley et al., 2004). As such, we will see that we can use this connection to better characterize the combinatorial aspects of ReLU optimization.

Our mathematical tools for describing the combinatorial structure of shallow ReLU network optimization include oriented hyperplane arrangements, polyhedral sets, polyhedral complexes, and zonotopes. Appendix A provides an approachable overview of these topics for unfamiliar readers.

Single Hidden Unit We start by considering a single ReLU unit $f_{\mathbf{w}}(\mathbf{x}) = \phi(\mathbf{w}^T\bar{\mathbf{x}})$ parameterized by the vector $\mathbf{w} \in \mathbb{R}^{d+1}$. Looking at its behavior as $\mathbf{w}$ ranges over $\mathbb{R}^{d+1}$ on a single training example $\mathbf{x}_i$, we see that there are two linear regimes depending on the sign of $\mathbf{w}^T\bar{\mathbf{x}}$. They are separated by the hyperplane in parameter space satisfying $\mathbf{w}^T\bar{\mathbf{x}} = 0$. We can describe such behavior mathematically as an oriented hyperplane with the sign of $\mathbf{w}^T\bar{\mathbf{x}}$ providing its orientation. The

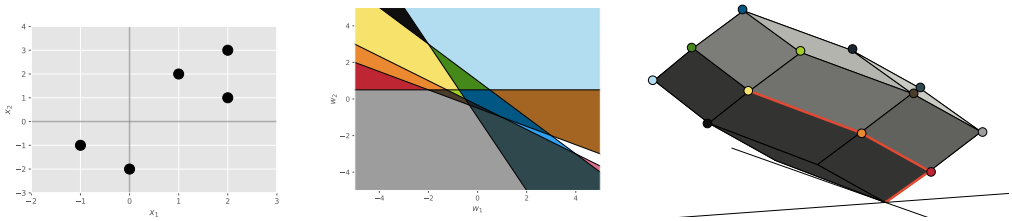

Figure 1: Left: A set of 5 training examples in $\mathbb{R}^2$. Center: A two dimensional slice of parameter space $\mathbb{R}^3$ along the $w_3 = 1$ plane reflecting the polyhedral complex $\mathcal{R}$. The lines correspond to the set of hyperplanes $H_1^0, \ldots, H_5^0$. The different shaded chambers correspond to different activation regions. Each chamber can be thought of as the base of cone whose apex is the origin. Right: The zonotope $\mathcal{Z}$ for this dataset. The corresponding activation region for a vertex is indicated by the colored circles. Note how the edges and faces of $\mathcal{Z}$ capture the incidence structure of the activation regions. Each of the red lines is a translation of a (homogenized) training example. Exactly these 3 training examples are active in the yellow activation region.

collection of oriented hyperplanes associated to each training example $\{\bar{\mathbf{x}}_i\}_{i=1}^N$ is known as an oriented hyperplane arrangement (Richter-Gebert & Ziegler, 2017).

Algorithm 1 Exact ERM (Arora et al., 2016)

Input: data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, 2nd layer $\mathbf{v} \in \mathbb{R}^{m+1}$
$\mathcal{A} \subseteq \{0,1\}^{m \times N}$ {feasible activation patterns (2)}
$W^* \in \mathbb{R}^{m \times (d+1)}$ {random initialization}
for $A \in \mathcal{A}$ **do**
  $W \leftarrow$ solution of (3)
  if $L(\tilde{W}) < L(W^*)$ **then**
    $W^* \leftarrow W$
  end if
end for
return W^*

The structure imposed on parameter space $\mathbb{R}^{d+1}$ by this oriented hyperplane arrangement can be described as a polyhedral complex (Ziegler, 2012), which is a collection of polyhedral sets and their faces that fit together in a "nice" way. The polyhedral complex $\mathcal{R}$ induced by the training set will contain codimension 0 sets called chambers. These correspond exactly to activation regions. Activation patterns have a one-to-one correspondence with the tuple of hyperplane orientations associated to each chamber. The center panel of fig. 1 provides an illustration of $\mathcal{R}$ for an example dataset.

The dual zonotope of a polyhedral complex is a single polytope providing an alternate representation of its combinatorial structure (Ziegler, 2012). Each dimension k member of the polyhedral complex has a corresponding codimension k face in the dual zonotope. Incidence relations between the members of the complex are preserved in the dual zonotope. Generally, a zonotope can be described as the image of an N-dimensional hypercube under a linear map whose columns are known as its generators (McMullen, 1971). Each vertex of the zonotope thus is a weighted sum of its generators with coefficients belonging to $\{0, 1\}$.

The dual zonotope $\mathcal{Z}$ of our polyhedral complex $\mathcal{R}$ has the training examples $\{\bar{\mathbf{x}}_i\}_{i=1}^N$ as its generators. The vertices of $\mathcal{Z}$ have a one-to-one correspondence to the activation regions of our network. When a vertex is expressed as a weighted sum over the generators, the coefficient $\{0, 1\}$ of each generator equals its corresponding example's value in the region's activation pattern. The right panel of fig. 1 shows an example zonotope $\mathcal{Z}$ and its duality with the polyhedral complex $\mathcal{R}$.

These correspondences allow us to assign additional structure to the set of activation patterns $\mathcal{A}$ rather than just treating it as an unstructured set. For example, the 1-skeleton of the zonotope $\mathcal{Z}$, which is the graph formed by its vertices and edges, provides a means of traversing the set of activation patterns. Furthermore, we can directly make connections between the training dataset and the activation pattern structure by making use of the fact that $\mathcal{Z}$ is generated by the training examples.

Multiple Hidden Units In the multiple hidden unit setting, i.e. $m > 1$, note that parameter space becomes an m-fold Cartesian product of single-unit parameter spaces. Furthermore, we are free to set the activation pattern for each unit independently of the others. As the combinatorial structure for

each hidden unit can be described using the zonotope $\mathcal{Z}$, the combinatorial structure for a multiple hidden unit network is described by the m-fold Cartesian product $\mathcal{Z}^m = \prod_{i=1}^m \mathcal{Z}$. As noted in appendix A.3.1, $\mathcal{Z}^m$ is also a zonotope. Each vertex of $\mathcal{Z}^m$ corresponds to a product of m vertices of $\mathcal{Z}$. As in the single unit case, there is a one-to-one correspondence between the vertices of $\mathcal{Z}^m$ and the set of activation patterns $\mathcal{A}$.

4 ReLU Optimization

4.1 NP-Hardness

Given the additional structure we have imposed on the set of activation patterns in algorithm 1, it is natural to ask whether we can use it to develop a global optimization algorithm that is more efficient than a brute-force search over activation patterns. Unfortunately, several works (Goel et al., 2020; Froese et al., 2021) have demonstrated that global optimization of a shallow ReLU network is NP-hard. Nevertheless, this does not preclude the existence of an efficient combinatorial optimizer given certain conditions on the input dataset. Since the zonotope $\mathcal{Z}^m$ encapsulates the combinatorial structure of the optimization problem, we look to see if properties of $\mathcal{Z}^m$ can be related to the difficulty of combinatorial optimization.

Nontrivial affine dependencies between training examples influence the combinatorial structure of the $\mathcal{Z}^m$. Since the reductions of NP-hard problems to ReLU optimization done in Goel et al. (2020); Froese et al. (2021) create datasets with such nontrivial dependencies, it is natural to ask whether it is NP-hard to optimize a shallow ReLU network over a training dataset in general position.

4.1.1 Discontinuity of the Global Optimum

If the global minimum of the loss is always continuous with respect to the input dataset, then the NP-hardness of optimization over arbitrary datasets in general position would follow from continuity since every set of points is arbitrarily close to a set in general position. However, we can prove that such continuity holds unconditionally for ReLU optimization only in the case where the training dataset is in general position. We give a sketch of the proof here along with some analysis of the failure cases that can happen when the data are not in general positions. We provide a full proof in appendix C.

Theorem 4.1. *Suppose we are given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ in general position and some $m \in \mathbb{N}$. Let $L^*(\mathcal{D})$ denote the global minimum of the loss* (1) *over the dataset $\mathcal{D}$ for a shallow ReLU network with m units. Given any $\epsilon > 0$, some $\delta > 0$ exists such $|L^*(\mathcal{D}) - L^*(\mathcal{D}_\epsilon)| < \delta$ for any dataset $\mathcal{D}_\epsilon = \{(\mathbf{x}'_i, y_i)\}_{i=1}^N$ satisfying $\|\mathbf{x}_i - \mathbf{x}'_i\|_2 \leq \epsilon$.*

Proof sketch. For a small enough perturbation, we can prove that the datasets' zonotopes are combinatorially equivalent. Hence their sets of feasible activation patterns will be exactly the same. Using the fact that any subset of a set in general position is also in general position, we can then show that the constrained convex optimization problem associated with each vertex is continuous with respect to the input dataset. Since the global minimum of the loss is just the minimum of the optimal loss for each vertex, its continuity follows from the fact that the composition of two continuous functions is continuous. □

When the dataset $\mathcal{D}$ is not in general position, there are two possible ways in which breaking of nontrivial affine dependencies between examples can cause the global minimum of the loss to become discontinuous. The first is that the globally optimal vertex in the perturbed zonotope exists in the original zonotope, but its associated constrained convex optimization problem is discontinuous with respect to the dataset. This can happen when there are nontrivial affine dependencies that get broken amongst the active examples in the vertex. The second way is that the globally optimal vertex of the perturbed zonotope does not exist in the original zonotope. Geometrically, we can think of such a vertex as resulting from the breakdown of a non-parallelepiped higher dimension face (Gover, 2014). See appendix D for examples of these phenomena.

Analysis of Reductions We can use this characterization of the instabilities of the global optimum to perturbations in the training data to analyze the reductions of NP-hard problems used in Goel et al. (2020). We focus on the reduction of the NP-hard set cover problem to the optimization of

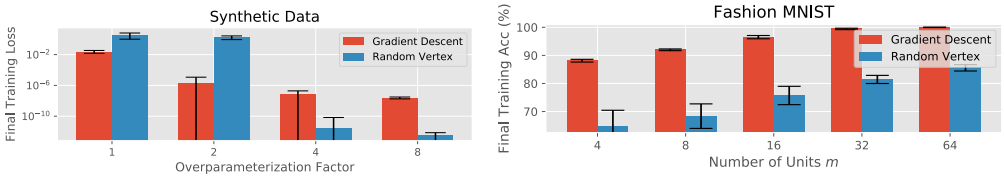

Figure 2: Comparison between gradient descent and optimization with a fixed random activation pattern. Left: Results for MSE on synthetic data for $d = 8$ and $m_{\text{gen}} = 8$. The overparameterization factor times m_{gen} equals the number of units in the trained network. Right: Results for accuracy on Fashion MNIST coat/pullover binary classification for $d = 16$ and $N = 700$.

a single bias-free ReLU. In appendix E.2, we provide an explicit example of an arbitrarily small perturbation that results in the global minimum of the loss being independent of the solution to the set cover problem.

To the best of our knowledge, existing reductions of NP-hard problems to ReLU optimization all create datasets that are not in general position (Goel et al., 2020; Froese et al., 2021). Therefore, we present a modification of the set cover reduction that produces a dataset in general position. See appendix E.3 for details of this modification along with a proof that it is indeed a reduction of the set cover problem. We thus have the following statement.

Theorem 4.2. *Optimizing a ReLU is NP-hard even when restricted to datasets in general position.*

4.2 Polynomial Time Optimization via Overparameterization

Even though ReLU network optimization is NP-hard in general, it can be shown that overparameterization allows for gradient descent to converge to the global minimum in polynomial time (Du et al., 2018; Zou & Gu, 2019; Oymak & Soltanolkotabi, 2020; Allen-Zhu et al., 2019). This is not a contradiction since optimization of overparameterized ReLU networks is a strict subset of the set of all ReLU optimization problems.

The general proof method of these works usually involves demonstrating that overparameterization results in activation patterns not changing much throughout training. This allows gradient descent to effectively bypass the combinatorial search of the outer loop in algorithm 1. The remaining optimization problem can then be shown to be similar to the constrained convex optimization problem (3) by assuming that the second layer is frozen. Using the zonotope formalism, we can interpret these results as saying that a sufficiently large number of hidden units m guarantees with high probability that a randomly chosen vertex corresponds to a region of parameter space containing a global minimum of the loss. Parameter initialization selects the random vertex in practice.

This can be justified theoretically through a connection to random feature models. Here we assume that the first layer is frozen, and the second layer forms a linear model over the random first layer features. As the number of units m increases past the number of training examples N, the set of first layer activations can become linearly independent. The probability of this approaches 1 as $m \to \infty$. Whether all of the parameters within an activation region produce linearly dependent activations can be shown to depend solely on its activation pattern when the dataset is in general position.

To test this, we ran experiments comparing batch gradient descent to solving (3) for a randomly chosen vertex on some toy datasets. We created synthetic datasets by first choosing the input dimension d and a positive integer m_{gen}. To get the training examples, we sampled $N = (d + 1)m_{\text{gen}}$ points in $\mathbb{R}^d$ i.i.d. from the standard Gaussian distribution. We then sampled the weights of a shallow ReLU network with m_{gen} units i.i.d. from the standard Gaussian distribution. We used this network to create the labels for our synthetic dataset. See appendix H.1.1 for details on the data generation process.

We also created toy binary classification datasets from MNIST (LeCun et al., 2010) and Fashion MNIST (Xiao et al., 2017) by choosing two classes, 5/9 and coat/pullover, respectively, to differentiate. We used the first $d \in \{8, 16\}$ components of the PCA whitened data and selected $N \in \{350, 700\}$ examples for our training sets. See appendix H.2 for details.

We present some of our results in fig. 2. See appendix H for details of the training procedures and for results on more d, m_{gen} and d, N pairs. On synthetic data, we see that the random vertex method finds a good solution for overparameterization factors of 4 and up. However, gradient descent tends

to arrive at reasonably good solutions for lower levels of overparameterization while the random vertex method fails. This was a general trend that we observed across different d, m_{gen} pairs on the synthetic datasets and d, N pairs on the binary classification datasets. Note that the Fashion MNIST networks represented in fig. 2 were relatively underparameterized with the maximal size of 64 units being overparameterized by only a factor of about 1.5.

This demonstrates that gradient descent can perform some aspects of the combinatorial search over zonotope vertices. We hypothesize that the gradient tends to be smaller within activation regions with a good optimum and thus gradient descent is more likely to stay within a good activation region. Conversely, the larger gradients within activation regions with poor optima make it more likely that a gradient descent step will move the parameters out of those regions. We can thus think of gradient descent as performing a pseudo-annealing process over the vertices of the zonotope since the likelihood of moving from one vertex to another decreases as the parameters settle into better activation regions.

4.2.1 Tighter Bounds

We now introduce a novel vertex selection scheme that runs in polynomial time and requires minimal overparameterization. Suppose $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is a dataset in general position. Assume that the examples are ordered by the value of their last coordinate, which we suppose is unique WLOG (i.e. $\mathbf{e}_d^T \mathbf{x}_i < \mathbf{e}_d^T \mathbf{x}_j$ for $i < j$). If not provided in this format, this can be accomplished in $O(N \log N)$ time. We now split the dataset into $\lceil \frac{N}{d+1} \rceil$ chunks containing at most $d+1$ examples. We write each chunk as $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=(k-1)(d+1)+1}^{\min(N, k(d+1))}$. Since each $\mathcal{D}_k$ contains a contiguous chunk of examples sorted along an axis in coordinate space, we see that we can always find a hyperplane separating $\mathcal{D}_k$ and $\mathcal{D}_{k'}$ for $k \neq k'$. For each $k = 1, \ldots, \lceil \frac{N}{d+1} \rceil$, we add two units to our ReLU network and assign them the activation pattern of 0 for examples belong to a $\mathcal{D}_{k'}$ with $k' < k$ and 1 for the remaining examples. One of the units will be multiplied by +1 in the second layer while the other will be multiplied by -1. Hence the network contains a total of $2\lceil \frac{N}{d+1} \rceil$ hidden units. We prove in appendix F that a set of weights with that activation pattern exists such that the output of the network on training examples exactly matches their labels. The key idea in the proof is that we can sequentially fit the examples in the k-th chunk without undoing our progress in fitting the chunks before it.

Theorem 4.3. *Given a dataset in $\mathbb{R}^d$ containing N examples in general position, a shallow ReLU network containing $2\lceil \frac{N}{d+1} \rceil$ hidden units can be found in polynomial time exactly fitting the dataset.*

To the best of our knowledge, this is the tightest known bound on the amount of overparameterization needed to find the global optimum of a ReLU network in polynomial time. A simple argument comparing the number of unknowns and the number of equations demonstrates that we need at least $\frac{N}{d+1}$ hidden units to exactly fit an arbitrary dataset with a shallow ReLU network. Hence our method uses only about twice as many hidden units as is necessary to fit the data. However, we emphasize that this ReLU optimization scheme is primarily of theoretical interest since we find in practice that the resulting ReLU network tends to be a very ill-conditioned function.

4.3 Relevance to Optimization in Practice

Practically all optimization of ReLU networks in practice uses some variant of gradient descent with an overparameterized network. As the degree of overparameterization goes down, gradient descent begins to arrive at increasingly suboptimal solutions (Nakkiran et al., 2021).

In section 4.2, we hypothesized how gradient descent can find activation regions containing good optima. However, the gradient of the loss is inherently a local property in parameter space while the space's decomposition into activation regions is inherently global. Boundaries between regions correspond to discontinuities in the gradient of the loss. We hypothesize that these properties lead to little direct information about the optimization problem being used to inform gradient descent's traversal over zonotope vertices. Hence we suspect that algorithms that explicitly traverse zonotope vertices using some loss-based criteria can outperform gradient descent in the underparameterized- to mildly-overparameterized regimes.

Algorithm 2 Greedy Local Search (GLS) Heuristic

Input: data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, 2nd layer $\mathbf{v} \in \mathbb{R}^{m+1}$, max steps $T \in \mathbb{N}$
$A_0 \in \text{vert}(\mathcal{Z}^m)$
for $t \in \{0, \ldots, T\}$ **do**
  $A_{t+1} \leftarrow A_t$
  for $A' \in \text{neighbors}(A_t)$ **do**
    if $\mathcal{L}^*(A'; \mathcal{D}) < \mathcal{L}^*(A_{t+1}; \mathcal{D})$ **then**
      $A_{t+1} \leftarrow A'$
    end if
  end for
  if $A_{t+1} = A_t$ **then**
    return A_t
  end if
end for
return A_T

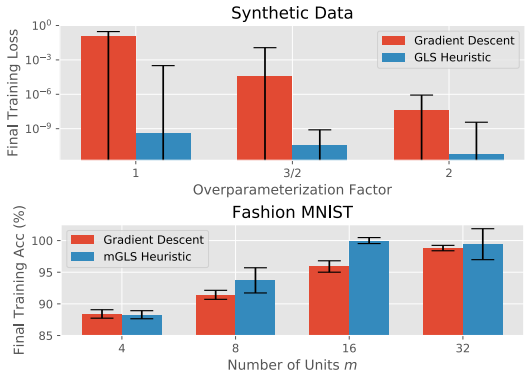

Figure 3: Comparison between gradient descent and our GLS heuristics. Top: Results for MSE on synthetic data for $d = 4$ and $m_{\text{gen}} = 2$. The overparameterization factor times m_{gen} equals the number of units in the trained network. Bottom: Results for accuracy on Fashion MNIST coat/pullover binary classification for $d = 8$ and $N = 350$.

4.3.1 Difficulty of Combinatorial Search

Unless $P = NP$, we are unlikely to find an efficient algorithm to perform the combinatorial search in algorithm 1 for arbitrary datasets (Goel et al., 2020). However, this does not preclude the existence of heuristics that tend to work well on problems encountered in practice. We investigated this by using a greedy local search (GLS) over the graph formed by the zonotope's 1-skeleton. We start by selecting a vertex at random and find its corresponding optimal loss by solving a convex program. We iterate over its neighboring vertices and compute their optimal losses as well. We then move to the neighboring vertex with the lowest loss and repeat the process until we arrive at a vertex with lower loss than its neighbors. We then take that vertex's optimal parameters as our approximation to the global minimization problem. This algorithm is defined in detail in algorithm 2.

We also experimented with some additional heuristics that help the GLS converge faster by reducing the number of convex problems solved at each step. For example, we can greedily move to the first neighboring vertex encountered with a lower loss, which significantly decreases the time per step in the early stages of training. We can further improve this by using geometric information about a solution's relative location in its activation region to try certain vertices first. We call the algorithm with these heuristics modified greedy local search (mGLS) and define it in detail in appendix G.

Note that these heuristics are not guaranteed to return a global minimizer of the loss. Furthermore, there are an exponential number of vertices in the zonotope, so there are no immediate guarantees of them taking less than exponential time to run. However, each step takes polynomial time since each vertex has $O(mN)$ neighbors, so each step solves a polynomial number of convex optimization problems.

We ran experiments comparing these heuristics to gradient descent on toy datasets generated in the same way as in section 4.2. We used GLS on the synthetic data and mGLS on the MNIST and Fashion MNIST derived data. We present some of our results in fig. 3. See appendix H for details of the training procedures and for results for more values of d, m_{gen} and d, N pairs. On synthetic data, the GLS heuristics significantly outperformed gradient descent. On binary classification tasks, mGLS outperformed gradient descent for networks with moderate levels of underparameterization and performed similarly otherwise. We observed a similar trend across the rest of the d, m_{gen} and d, N pairs. These results support our hypothesis that gradient descent is suboptimal in the combinatorial search over activation patterns. Furthermore, they suggest that this combinatorial optimization might tend to be tractable in practice.

5 Related Work

Some of the concepts in this work also arise in Zhang et al. (2018). A key difference is that they analyze the activation regions in input space given a ReLU network with fixed parameters. We can, in fact, use their tropical geometric approach to derive our zonotope formalism for a single ReLU unit. To do so, the roles of the weights and data must be swapped; we instead use a fixed data matrix and varying vector of weights while they use a fixed weight matrix and varying vector of data. Our use of a zonotope generated by the training examples, however, is novel. Misiakos et al. (2021) show that the approximation error between two shallow ReLU networks depends on the the Hausdorff distance between the zonotopes generated by the each network's units. Bach (2017) also use a Hausdorff distance between zonotopes in the context of neural network optimization.

Goel et al. (2020) provide proofs of the NP-hardness of optimization of shallow ReLU networks and the hardness of even finding an approximate solution. Froese et al. (2021) extend these results and show that the brute force search in Arora et al. (2016) cannot be avoided in the worst case. Du et al. (2018) was one of the first works to prove that overparameterization in shallow ReLU networks allows gradient descent to converge to a global optimum in polynomial time. Their bound of $\Omega(N^6)$ on the number of hidden units needed for convergence was improved upon by subsequent work (Ji & Telgarsky, 2019; Daniely, 2019). For example, Oymak & Soltanolkotabi (2020) proved a bound of $\Omega(N^2/d)$.

Pilanci & Ergen (2020) and Wang et al. (2021) represent global optimization of shallow ReLU networks with ℓ_2 regularization using a convex optimization problem that operates simultaneously over *all* activation patterns for a single unit. Multiple units are handled by summing over the activations with different activation patterns. This leads to exponential complexity in the data dimension d but avoids exponential complexity in the number of units m. Dey et al. (2020) provide an example of a heuristic algorithm that searches over activation patterns for a single ReLU unit. Their algorithm operates on the principle that examples with large positive labels are more likely to belong to the active set in good solutions.

6 Conclusion

We introduced a novel characterization of the combinatorial structure of activation patterns implicit in the optimization of shallow ReLU networks. We showed that it can be described as a Cartesian product of zonotopes generated by the training examples. We used this zonotope formalism to explore aspects of the optimization of shallow ReLU networks. It provides a natural way to describe instabilities of the global minimum to perturbations of the dataset. We then related this to work on the NP-hardness of global ReLU optimization. In particular, we demonstrated that this optimization problem is still NP-hard even when restricted to datasets in general position, which is commonly assumed of data in practice.

We then explored how combinatorial considerations play into the relationship between overparameterization and polynomial-time optimization of shallow ReLU networks. Namely we interpret known results for gradient descent as stating that a randomly chosen zonotope vertex will be close to one whose activation region contains a good local optimum. We then provide empirical evidence that sufficient overparameterization makes it highly likely that a randomly chosen vertex has a good local optimum. We also provide a polynomial-time algorithm that can find a vertex containing the global optimum using approximately twice the minimum number of hidden units needed to fit the dataset exactly. Finally, we provide a GLS heuristic over zonotope vertices that outperforms gradient descent on some toy problems.

In future work we plan to theoretically and empirically explore heuristics and algorithms that perform well on real-world datasets. We hope to analyze how vertex choice impacts generalization. Further insights might be derived by exploring the connections of hyperplane arrangements to tropical geometry and oriented matroids (Stanley et al., 2004; Oxley, 2006; Maclagan & Sturmfels, 2015). One caveat of our theory is that it applies to only a shallow ReLU network. However, the concepts of activation patterns are still meaningful for deep ReLU networks but require real algebraic geometry for analysis (Basu, 2014; Bochnak et al., 2013). We hope that further research along these avenues will deepen our understanding of neural network training and enable improvements to training in practice.

References

Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., and Kolter, Z. Differentiable convex optimization layers. *arXiv preprint arXiv:1910.12430*, 2019.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Basu, S. Algorithms in real algebraic geometry: a survey. *arXiv preprint arXiv:1409.1534*, 2014.

Bochnak, J., Coste, M., and Roy, M.-F. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

Daniely, A. Neural networks learning and memorization with (almost) no over-parameterization. *arXiv preprint arXiv:1911.09873*, 2019.

Dey, S. S., Wang, G., and Xie, Y. Approximation algorithms for training one-node relu neural networks. *IEEE Transactions on Signal Processing*, 68:6696–6706, 2020.

Diamond, S. and Boyd, S. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

Domahidi, A., Chu, E., and Boyd, S. Ecos: An socp solver for embedded systems. In *2013 European Control Conference (ECC)*, pp. 3071–3076. IEEE, 2013.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Froese, V., Hertrich, C., and Niedermeier, R. The computational complexity of relu network training parameterized by data dimensionality. *arXiv preprint arXiv:2105.08675*, 2021.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Goel, S., Klivans, A., Manurangsi, P., and Reichman, D. Tight hardness results for training depth-2 relu networks. *arXiv preprint arXiv:2011.13550*, 2020.

Gover, E. Congruence and metrical invariants of zonotopes. *arXiv preprint arXiv:1401.4749*, 2014.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2): 251–257, 1991.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Maclagan, D. and Sturmfels, B. *Introduction to tropical geometry*, volume 161. American Mathematical Soc., 2015.

McMullen, P. On zonotopes. *Transactions of the American Mathematical Society*, 159:91–109, 1971.

Misiakos, P., Smyrnis, G., Retsinas, G., and Maragos, P. Neural network approximation based on hausdorff distance of tropical zonotopes. In *International Conference on Learning Representations*, 2021.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Oxley, J. G. *Matroid theory*, volume 3. Oxford University Press, USA, 2006.

Oymak, S. and Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Pilanci, M. and Ergen, T. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pp. 7695–7705. PMLR, 2020.

Richter-Gebert, J. and Ziegler, G. M. *6: Oriented matroids*. Chapman and Hall/CRC, 2017.

Stanley, R. P. et al. An introduction to hyperplane arrangements. *Geometric combinatorics*, 13 (389-496):24, 2004.

Wang, Y., Lacotte, J., and Pilanci, M. The hidden convex optimization landscape of regularized two-layer relu networks: an exact characterization of optimal solutions. In *International Conference on Learning Representations*, 2021.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Zhang, L., Naitzat, G., and Lim, L.-H. Tropical geometry of deep neural networks. In *International Conference on Machine Learning*, pp. 5824–5832. PMLR, 2018.

Ziegler, G. M. *Lectures on polytopes*, volume 152. Springer Science & Business Media, 2012.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. *arXiv preprint arXiv:1906.04688*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our contributions are primarily theoretical.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide proofs in the appendix.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have included this in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We include standard deviations for all results in the tables.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the datasets and software libraries used.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The datasets used (MNIST, Fashion MNIST) are standard in machine learning.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] They are widely used datasets with pictures of handwritten digits and items of clothing.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Mathematical Background

A.1 Oriented Hyperplane Arrangements

Any $\bar{\mathbf{x}} \in \mathbb{R}^{d+1}$ defines a tripartite division of $\mathbb{R}^{d+1}$ given by

$$H^{\alpha} = \{\mathbf{w} \in \mathbb{R}^{d+1} \mid \operatorname{sign}(\mathbf{w}^T \bar{\mathbf{x}}) = \alpha\}, \tag{4}$$

where $\alpha \in \{-1, 0, +1\}$. The set H^0 is a hyperplane while H^+, H^- are the positive and negative open half-spaces, respectively. We call their closures $\bar{H}^{\pm} = H^0 \cup H^{\pm}$ the positive/negative closed half-spaces. For consistency, we say $\bar{H}^0 = H^0$.

We are provided with a finite set of vectors $\{\bar{\mathbf{x}}_i\}_{i=1}^N \subseteq \mathbb{R}^{d+1}$. We interpret these vectors as the training examples in section 3, but for now consider them to be arbitrary vectors. Let H_i^{α} denote the subsets given by (4) for the i-th example. The union of the hyperplanes $\bigcup_{i=1}^N H_i^0$ separates chambers of $\mathbb{R}^{d+1}$ into a finite number of disjoint cells. Each cell can be described as the intersection of open half-spaces and thus can be indexed by a vector of sign patterns reflecting whether the positive or negative half-space is used. The structure induced by these hyperplanes along with their orientation information creates what is known as an oriented hyperplane arrangement (Richter-Gebert & Ziegler, 2017).

A.2 Polyhedral Complexes

We can also describe this arrangement through the notion of a polyhedral complex (Ziegler, 2012). We first provide some definitions that will be useful. A polyhedral set is any set that is the intersection of a finite number of closed half-spaces. A polytope is a bounded polyhedral set. Any hyperplane intersecting a polyhedral set will either divide it into two polyhedral sets or only intersect it on its boundary. In the latter case, we call such a hyperplane a supporting hyperplane. A face of a polyhedral set is defined as its intersection with a supporting hyperplane. The dimension of a face is the dimension of its affine span. If the dimension of a face is k, we call it a k-face. We call 0-faces vertices and 1-faces edges. By convention, the empty set is a face of any polyhedral set. The 1-skeleton of a polyhedral set is the graph formed by its vertices and edges.

A polyhedral complex $\mathcal{K}$ is a finite set of polyhedral sets satisfying

1. If $P \in \mathcal{K}$ and F is a face of P, then $F \in \mathcal{K}$.
2. If $P_1, P_2 \in \mathcal{K}$, then their intersection $P_1 \cap P_2$ is a face of both P_1, P_2.

We call a codimension 0 member of $\mathcal{K}$ a chamber of the polyhedral complex. The support of a polyhedral complex is the union of its polyhedral sets. If a polyhedral complex's support equals the entire space, then we call it a polyhedral decomposition of the space. The "is a face of" relation induces a poset structure on the members of a polyhedral complex, which we call its face poset. This can be extended further to a meet-semilattice with the meet operation being given by set intersection. We call this the face semilattice.

Going back to the case of an oriented hyperplane arrangement, let $\mathbf{a} \in \{-1, 0, +1\}^N$ be some sign pattern. Now let us write

$$R^{\mathbf{a}} = \bigcap_{i=1}^{N} \bar{H}_i^{a_i}. \tag{5}$$

Since our hyperplanes are all linear (i.e. non-affine), we always have $\mathbf{0} \in R^{\mathbf{a}}$. If $R^{\mathbf{a}} = \{\mathbf{0}\}$, we say that $R^{\mathbf{a}}$ is null. Define $\mathcal{R}$ to be the set of all $R^{\mathbf{a}}$. Then $\mathcal{R}$ is a polyhedral complex, which we prove in appendix B.1.

Every chamber of $\mathcal{R}$ corresponds to a non-null $R^{\mathbf{a}}$ with $\mathbf{a} \in \{-1, +1\}^N$. If the hyperplanes are general positions, this correspondence is one-to-one. The face semilattice of $\mathcal{R}$ provides information about how its chambers are arranged in space. For example, let $M \in \mathcal{R}$ be the meet of two chambers. If $M \neq \{\mathbf{0}\}$, then those chambers are neighbors. Then $\dim M \in \{1, \ldots, d\}$ and the sign patterns of the chambers differ by at least $d + 1 - \dim M$ sign flips, with equality always holding the hyperplanes are in general positions.

A.3 Dual Zonotopes

It turns out that we can describe the incidence structure of the polyhedral complex $\mathcal{R}$ nicely with a single polytope called its dual zonotope $\mathcal{Z}$ (Ziegler, 2012). A zonotope is any polytope that can be expressed as the Minkowski sum of a finite set of line segments called its generators (McMullen, 1971). In the case of the dual zonotope of $\mathcal{R}$, these generators are the line segments $\{[\mathbf{0}, \bar{\mathbf{x}}_i]\}_{i=1}^N$. We can thus write

$$\mathcal{Z} = \left\{ \sum_{i=1}^{N} \lambda_i \bar{\mathbf{x}}_i \mid \lambda_i \in [0,1] \right\}. \tag{6}$$

When the generators are in general positions, each k-face of $\mathcal{Z}$ is a k-dimension parallelepiped. Nontrivial linear dependencies between generators, however, lead to k-faces that are the union of multiple k-dimension parallelepipeds lying in the same k-dimension affine subspace.

The duality between $\mathcal{Z}$ and $\mathcal{R}$ allows us to associate members of $\mathcal{R}$ with faces of $\mathcal{Z}$. Each k-face of $\mathcal{Z}$ corresponds to a codimension k member of $\mathcal{R}$. Notably, the vertices of $\mathcal{Z}$, denoted by $\mathrm{vert}(\mathcal{Z})$, correspond to the chambers of $\mathcal{R}$. Relationships between members of $\mathcal{R}$ carry over to faces of $\mathcal{Z}$. For example, two neighboring chambers of $\mathcal{R}$ whose sign patterns differ by a single flipped sign will correspond to two vertices connected by an edge in $\mathcal{Z}$.

We now describe how to make this correspondence explicit. Let $\mathbf{v} \in \mathrm{vert}(\mathcal{Z})$ be a vertex. It can be shown that $\mathbf{v}$ has a unique representation as $\sum_{i=1}^N \lambda_i \bar{\mathbf{x}}_i$ with every $\lambda_i \in \{0,1\}$. We call the vector $(\lambda_1, \ldots, \lambda_N)$ the barcode of the vertex. We will often treat a vertex interchangeably with its barcode in this paper with difference being clear by context. Let $\mathbf{a}$ be the sign pattern of the chamber corresponding to $\mathbf{v}$. Then $a_i = -1$ if $\lambda_i = 0$ and $a_i = +1$ if $\lambda_i = 1$ for $i = 1, \ldots, N$.

Now suppose two vertices $\mathbf{v}_1, \mathbf{v}_2 \in \mathrm{vert}(\mathcal{Z})$ are connected by an edge, and that the hyperplanes of $\mathcal{R}$ are in general positions. Then there exists a single i^* such that, WLOG, $\mathbf{v}_2 = \mathbf{v}_1 + \bar{\mathbf{x}}_{i^*}$. The sign pattern of the member of $\mathcal{R}$ corresponding to the edge can then be found by finding the sign pattern for $\mathbf{v}_1$ and changing its i^*-th entry to be 0.

A.3.1 Cartesian Power of Zonotopes

Let us consider an m-ary Cartesian power of a zonotope $\mathcal{Z}^m = \prod_{i=1}^m \mathcal{Z}$. We can see that $\mathcal{Z}^m$ is also a zonotope and is generated by line-segments from the origin to members of $\bigcup_{i=1}^m \bigcup_{j=1}^N \{\mathbf{e}_i \bar{\mathbf{x}}_j^T\}$, where $\mathbf{e}_i \in \mathbb{R}^m$ is the i-th standard coordinate vector. Each k-face of $\mathcal{Z}^m$ is the Cartesian product of a set of $\{k_1, \ldots, k_m\}$-faces of $\mathcal{Z}$ where $k = k_1 + \cdots + k_m$. Notably, each vertex of $\mathcal{Z}^m$ corresponds to a product of m vertices of $\mathcal{Z}$. Edges of $\mathcal{Z}^m$ correspond to the product of a single edge of $\mathcal{Z}$ with $m-1$ vertices.

B Proofs for Appendix A

B.1 Proof that $\mathcal{R}$ is a Polyhedral Complex

Recall that a polyhedral complex $\mathcal{K}$ is a finite set of polyhedral sets satisfying

1. If $P \in \mathcal{K}$ and F is a face of P, then $F \in \mathcal{K}$.
2. If $P_1, P_2 \in \mathcal{K}$, then their intersection $P_1 \cap P_2$ is a face of both P_1, P_2.

Recall that we have defined $\mathcal{R}$ as

$$\mathcal{R} = \left\{ R^{\mathbf{a}} \subseteq \mathbb{R}^{d+1} \mid \mathbf{a} \in \{-1, 0, +1\}^N \right\}, \tag{7}$$

where $R^{\mathbf{a}}$ is given by (5). Note that generally $|\mathcal{R}| \leq 3^N$ since multiple $R^{\mathbf{a}}$ can equal $\{\mathbf{0}\}$.

It is easy to see that every $R^{\mathbf{a}} \in \mathcal{R}$ is a polyhedral set since it can be can defined as the intersection of finitely many closed half-spaces. When $a_i = 0$, its corresponding hyperplane in (5) is equivalent to the intersection of its positive and negative closed half-spaces.

We now prove the first condition for $\mathcal{R}$ being a polyhedral complex. Suppose $R^{\mathbf{a}} \in \mathcal{R}$ and suppose F is a face of $R^{\mathbf{a}}$. Hence F is the intersection of $R^{\mathbf{a}}$ with a supporting hyperplane. It is straightforward

to see that any face of $R^{\mathbf{a}}$ can be represented by a $R^{\mathbf{b}}$ where $b_i = a_i$ for all $i \in [N] \setminus I$ and $b_i = 0, a_i = \pm 1$ for $i \in I \subseteq [N]$. Hence the face $F = R^{\mathbf{b}} \in \mathcal{R}$.

We now prove the second condition for $\mathcal{R}$ being a polyhedral complex. Let $R^{\mathbf{a}}, R^{\mathbf{b}} \in \mathcal{R}$. From (5), we see that

$$R^{\mathbf{a}} \cap R^{\mathbf{b}} = \bigcap_{i=1}^{N} \bar{H}_i^{a_i} \cap \bar{H}_i^{b_i}. \tag{8}$$

We see that $H_i^{a_i} \cap \bar{H}_i^{b_i} = H_i^{\pm 1}$ if $H_i^{a_i} = H_i^{b_i} = H_i^{\pm 1}$, and that $H_i^{a_i} \cap \bar{H}_i^{b_i} = H_i^0$ otherwise. It is easy to see that this intersection can be represented as the intersection of a supporting hyperplane with either $R^{\mathbf{a}}$ or $R^{\mathbf{b}}$. Hence their intersection is a mutual face.

C Proof of Theorem 4.1

Suppose we are given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ in general position. Given some $\epsilon > 0$, let $\mathcal{D}_\epsilon = \{(\mathbf{x}'_i, y_i)\}_{i=1}^N$ be any perturbation of $\mathcal{D}$ such that $\|\mathbf{x}_i - \mathbf{x}'_i\|_2 \leq \epsilon$.

Let $\bar{X} \in \mathbb{R}^{N \times (d+1)}$ denote the data matrix in homogeneous coordinates for $\mathcal{D}$. Let $\bar{X}'$ denote the corresponding data matrix for $\mathcal{D}_\epsilon$. We see that $\|\bar{X} - \bar{X}'\|_F \leq \sqrt{N}\epsilon$. Hence some matrix $P \in \mathbb{R}^{N \times (d+1)}$ exists such that $\bar{X}' = \bar{X} + P$ and $\|P\|_F \leq \sqrt{N}\epsilon$.

We can interpret the zonotope definition (6) as saying that a zonotope is the image of a hypercube under the matrix formed by its generators. Hence if $\mathcal{Z}$ is the zonotope of the original dataset, we may write

$$\mathcal{Z} = \left\{ \bar{X}^T \mathbf{u} \in \mathbb{R}^{d+1} \mid \mathbf{u} \in [0,1]^N \right\}. \tag{9}$$

Let $\mathcal{Z}'$ denote the corresponding zonotope for the perturbed dataset.

We now wish to show that the vertices of $\mathcal{Z}$ are in a one-to-one correspondence with the vertices $\mathcal{Z}'$ for sufficiently small ϵ with their sets of vertex barcodes coinciding. Let $\mathbf{b} \in \{0,1\}^N$ be some binary vector such that $\mathbf{p} = \bar{X}^T \mathbf{b}$ is a vertex of $\mathcal{Z}$. Then we know that some affine hyperplane $H \subseteq \mathbb{R}^{d+1}$ exists such that $H \cap \mathcal{Z} = \{\mathbf{p}\}$.

Let $\mathbf{q} = \bar{X}^T \mathbf{c}$, where $\mathbf{c} \in \{0,1\}^N$, be the image of an arbitrary vertex of the hypercube such that $\mathbf{q} \neq \mathbf{p}$. Note that $\mathbf{q}$ is not necessarily a vertex of $\mathcal{Z}$. We thus find some $\delta > 0$ such that the distance from $\mathbf{q}$ to the hyperplane H is greater than δ for every such $\mathbf{q}$. Furthermore, all such $\mathbf{q}$ will lie in exactly one of the half-spaces formed by the hyperplane.

Let $H' \subseteq \mathbb{R}^{d+1}$ be the affine hyperplane formed by shifting H by $P^T \mathbf{b}$. If we let $\mathbf{p}' = \bar{X}'^T \mathbf{b}$, it is straightforward to see that $\mathbf{p}' \in H'$. Hence H' intersects the perturbed zonotope $\mathcal{Z}'$.

Let $\mathbf{q}' = \bar{X}'^T \mathbf{c}$. Note that we can write $\mathbf{p}' = \mathbf{p} + P^T \mathbf{b}$ and $\mathbf{q}' = \mathbf{q} + P^T \mathbf{c}$. Let us now bound $\|P^T \mathbf{b}\|_2$. We see that $\|\mathbf{b}\|_2 \leq \sqrt{N}$. Using known relations between matrix norms, we see that $\|P^T\|_2 \leq \|P^T\|_F \leq \sqrt{N}\epsilon$. Hence $\|P^T \mathbf{b}\|_2 \leq N\epsilon$. By the exact same logic, we see that $\|P^T \mathbf{c}\|_2 \leq N\epsilon$.

Now suppose that we choose $\epsilon < \frac{\delta}{2N}$. By the triangle inequality, we can see that $\mathbf{q}'$ can move a distance at most $2N\epsilon < \delta$ relative to the hyperplane H'. Since the distance from $\mathbf{q}$ to H was greater than δ, we see that every such $\mathbf{q}'$ must lie on the same side of the hyperplane H'. Hence $H' \cap \mathcal{Z}' = \{\mathbf{p}'\}$, which implies that $\mathbf{p}'$ is a vertex of $\mathcal{Z}'$. Hence every vertex barcode of $\mathcal{Z}$ is a vertex barcode of $\mathcal{Z}'$. As $\mathcal{D}_\epsilon$ will also be in general position for small enough ϵ, we can swap the roles of $\mathcal{Z}$ and $\mathcal{Z}'$ in our proof to see that every vertex barcode of $\mathcal{Z}'$ is a vertex barcode of $\mathcal{Z}$. It thus follows that there is a one-to-one correspondence between vertex barcodes of $\mathcal{Z}$ and $\mathcal{Z}'$.

By the relationship between zonotope vertices and activation regions shown in section 3, we have thus proved that the set of convex optimization problems for $\mathcal{D}$ are a slightly perturbed version of the convex optimization problems for $\mathcal{D}_\epsilon$.

As any subset of a set of vectors in general position is also in general position, we can see that the solution of each convex optimization problem is continuous with respect to perturbations of the data (Agrawal et al., 2019). The global minimum of the loss is given by the minimum over the set of per-activation-region local minima. Hence the global minimum is continuous with respect to the training dataset as it is the composition of two continuous functions.

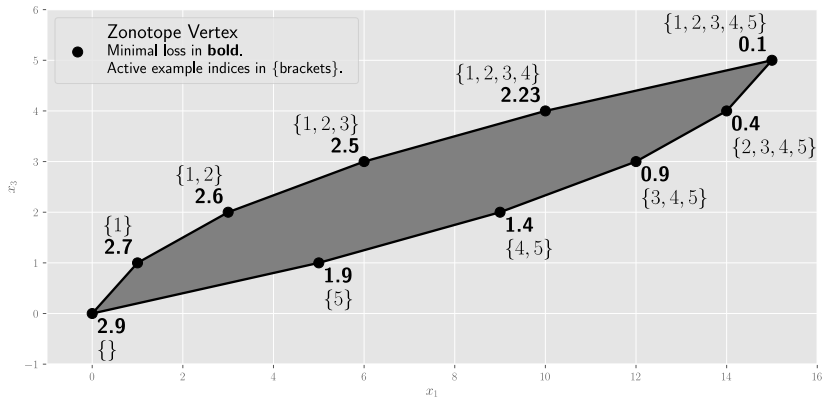

Figure 4: The zonotope $\mathcal{Z}$ associated to the dataset introduced in appendix D.1.

D Examples of Discontinuities at Datasets not in General Position

This section provides examples of the two sources of discontinuities of the minimal loss of a dataset when it is not in general position.

D.1 Convex Problem Associated to a Vertex being Discontinuous

Here we provide an example of a dataset whose optimal loss is not continuous with respect to the dataset. For this dataset, the optimal vertex is the same and exists in both the original and perturbed zonotopes. This means that its associated convex optimization problem is discontinuous with respect to the dataset.

Let us consider the dataset $\mathcal{D} \subseteq \mathbb{R}^2 \times \mathbb{R}$ given by

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
$$Y = \begin{bmatrix} 1 & 2 & 2.5 & 4 & 5 \end{bmatrix}.$$

Note that $\mathcal{X}$ lies entirely within the $x_2 = 0$ hyperplane and thus is not in general position.

Now consider the problem of optimizing a single affine ReLU over $\mathcal{D}$ with respect to the L1 loss. We assume a linear second layer and take the ReLU's corresponding second layer weight to be 1. The zonotope $\mathcal{Z}$ associated to this optimization problem is presented in fig. 4.[1] Each vertex has been labeled with its minimal loss in bold and with its set of active example indices. The vertex with the smallest loss of 0.1 is active on all of the examples.

Now consider what happens when we perform the following perturbation on the dataset

$$X_\epsilon = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & \epsilon & 0 & 0 \end{bmatrix},$$

where $\epsilon > 0$ is arbitrarily small. Let $\mathcal{D}_\epsilon = (X_\epsilon, Y)$. If we set the parameters $(\mathbf{w}, b)$ of our ReLU to $\mathbf{w} = \begin{bmatrix} 1 & -\frac{1}{2\epsilon} \end{bmatrix}^T$ and $b = 0$, we see that we fit $\mathcal{D}_\epsilon$ exactly and thus obtain zero L1 loss. These parameters belong to the same vertex as the global minimum of the unperturbed dataset. Hence we conclude that the convex optimization problem associated to this vertex is discontinuous with respect to the dataset.

D.2 Global Loss in New Vertex of Perturbed Zonotope

Here we provide an example of a dataset whose optimal loss is not continuous with respect to the dataset. For this dataset, the optimal vertex in the perturbed zonotope does not exist in the original zonotope.

[1] Technically this is a slice of the zonotope along the $x_2 = 0$ plane. The full zonotope is equal to the cylinder $\mathcal{Z} + \mathbb{R}\mathbf{e}_2$.

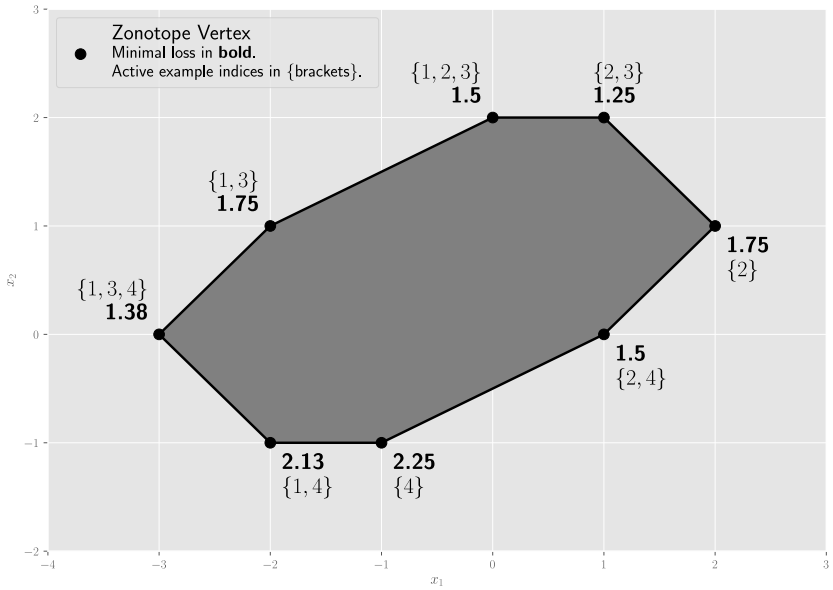

Figure 5: The zonotope $\mathcal{Z}$ associated to the dataset introduced in appendix D.2.

Let us consider the dataset $\mathcal{D} \subseteq \mathbb{R}^3 \times \mathbb{R}$ given by

$$X = \begin{bmatrix} -1 & 2 & -1 & -1 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$
$$Y = \begin{bmatrix} 4 & 3 & 2 & 1 \end{bmatrix}.$$

Consider the problem of optimizing a single linear ReLU over $\mathcal{D}$ with an L1 loss. Assume that the second layer is linear with the ReLU's output weight set to 1. Note that $\mathcal{X}$ lies entirely within the $x_3 = 0$ plane and thus is not in general position. The zonotope $\mathcal{Z}$ associated to this optimization problem is presented in fig. 5.[2] Each vertex has been labeled with its minimal loss in bold and with its set of active example indices. The vertex with the smallest loss of 1.25 has the examples $\{\mathbf{x}_2, \mathbf{x}_3\}$ active.

Now consider what happens when we perform the following perturbation on the dataset

$$X_\epsilon = \begin{bmatrix} -1 & 2 & -1 & -1 \\ 0 & 1 & 1 & -1 \\ 0 & \epsilon & 0 & 0 \end{bmatrix} \tag{10}$$

where $\epsilon > 0$ is arbitrarily small. Let $\mathcal{D}_\epsilon = (X_\epsilon, Y)$. The global minimum of the loss for $\mathcal{D}_\epsilon$ occurs at the parameter value $\mathbf{w} = \begin{bmatrix} -\frac{3}{2} & \frac{1}{2} & \frac{11}{2\epsilon} \end{bmatrix}^T$. The loss value at this point is 0.625, and these parameters are associated to the zonotope vertex with all examples active. As evident from fig. 5, such a vertex does not exist in the unperturbed zonotope.

E NP-Hardness

Goel et al. (2020) prove the NP-hardness of optimizing a single ReLU by reducing solving an instance of the NP-hard set cover problem to optimizing a single ReLU over a train dataset. In the set cover problem, we are given a collection $\mathcal{T} = \{T_1, \ldots, T_M\}$ of subsets of a given set U. Given some $t \in \mathbb{N}$, the goal is determine whether a subcollection $\mathcal{S} \subseteq \mathcal{T}$ exists such that $U = \bigcup_{S \in \mathcal{S}} S$ and $|\mathcal{S}| \leq t$.

[2]Technically this is a slice of the zonotope along the $x_3 = 0$ plane. The full zonotope is equal to the cylinder $\mathcal{Z} + \mathbb{R}\mathbf{e}_3$.

E.1 Reduction to ReLU Optimization

Goel et al. (2020) use a single ReLU without a bias as their model, so we may write our network as

$$f_{\mathbf{w}}(\mathbf{x}) = \phi(\mathbf{w}^T\mathbf{x}). \tag{11}$$

The input dimension of their model is $d = M + 2$. Of these dimensions, M correspond to members of $\mathcal{T}$ and two are used as "constraint coordinates". We use $\mathbf{e}_{T_i}$ to denote a unit coordinate vector corresponding to T_i, and $\mathbf{e}_\gamma$ and $\mathbf{e}_1$ as the unit coordinate vectors for the constraint coordinates.

Set $\gamma = 0.01/M^2$. Overall, they create $N = |U| + M + 2$ labeled training examples. For the constraint coordinates, they create the examples

$$(\mathbf{x}_\gamma, y_\gamma) = (\mathbf{e}_\gamma, \gamma) \tag{12}$$

and

$$(\mathbf{x}_1, y_1) = (\mathbf{e}_1, 1). \tag{13}$$

For each $T_i \in \mathcal{T}$, they create the example

$$(\mathbf{x}_{T_i}, y_{T_i}) = (\mathbf{e}_\gamma + \mathbf{e}_{T_i}, \gamma). \tag{14}$$

For each $u \in U$, they create the example

$$(\mathbf{x}_u, y_u) = (\mathbf{e}_1 + \sum_{T_i \ni u} \mathbf{e}_{T_i}, 0). \tag{15}$$

Let $\mathcal{D}$ denote the entire labeled dataset, and let $\mathcal{X}$ denote the just the examples without labels. Note that both $\mathcal{D}$ and $\mathcal{X}$ will generally be multisets since if $u, u' \in U$ belong to exactly the same set of subsets in $\mathcal{T}$, then $\mathbf{x}_u = \mathbf{x}_{u'}$.

Using mean squared error, the training loss of the network can be written as

$$L(\mathbf{w}) = \frac{1}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}} (f_{\mathbf{w}}(\mathbf{x}) - y)^2. \tag{16}$$

Goel et al. (2020) show that if the union of t or fewer members of $\mathcal{T}$ equals U, then the global minimum of this loss will be less than or equal to $t\gamma^2/N$. The weights $\mathbf{w} \in \mathbb{R}^d$ corresponding to this optima will have $w_\gamma = \gamma$, $w_1 = 1$, $w_{T_i} = -1$ for all $T_i \in \mathcal{S}$, and all other parameters set to zero.

The activation pattern for this optima will have $\mathbf{x}_1$ and $\mathbf{x}_\gamma$ being active while $\mathbf{x}_u$ is inactive for every $u \in U$. Any $\mathbf{x}_{T_i}$ will be inactive if $T_i \in \mathcal{S}$ and active otherwise.

E.2 Discontinuous Response to Perturbation

Here, we demonstrate how to create a perturbed dataset $\mathcal{D}'_\epsilon$ so that the the global minimum of loss will always be at most γ^2/N regardless of the solution to the set cover problem. This will hold true even as the scale of the perturbation $\epsilon > 0$ approaches 0.

In this perturbation, we pick an arbitrary $T \in \mathcal{T}$ and set $\mathbf{x}'_u = \mathbf{x}_u + \epsilon\mathbf{e}_T$ for the examples corresponding to all $u \in U$. All other examples are left unchanged.

Clearly, $\|\mathbf{x} - \mathbf{x}'\|_2 \leq \epsilon$ for all original-perturbed example pairs. To find a $\mathbf{w}$ with a loss value of γ^2/N, set $w_\gamma = \gamma$, $w_1 = 1$, $w_T \leq -\epsilon^{-1}$, and set all other parameters to 0. In this case, the model's predictions are correct for every training example except $(\mathbf{x}_T, y_T)$. In this case, the model predicts 0 while the label is γ, so the total loss is γ^2/N.

Note that the L2 norm of the parameters at the global minimum for the perturbed dataset approaches infinity as the size of the perturbation ϵ approaches 0.

The activation pattern at this optima will be the same as the activation pattern at the unperturbed optima except that $\mathbf{x}_T$ will be inactive while the rest of the $\mathbf{x}_{T_i}$ will be active.

We note that this activation pattern is achievable on the unperturbed dataset. However, it requires setting w_1 to a small positive value, w_γ to a relatively large positive value, all w_{T_i} for $T_i \neq T$ to moderate negative values, and w_T to a large enough negative value. For example, setting $w_1 = \gamma$, $w_\gamma = 2$, $w_{T_i} = -1$ for $T_i \neq T$, and $w_T = -3$ works. Hence as discussed in section 4.1.1, this discontinuity corresponds to a discontinuity in the constrained convex optimization problem (3) associated to a vertex rather than the new optimum occurring at a vertex not present in the original zonotope.

E.3 Reduction to Dataset in General Position (Proof of Theorem 4.2)

It is possible to perturb some of the examples in appendix E.1 to get a dataset in general position that is still the reduction of the subset-sum problem.

Theorem E.1. *Let $\delta_1, \delta_2 \in \mathbb{R}$ be constants satisfying $0 < \delta_1 < \delta_2 < \frac{1}{2d}$. For each $u \in U$, replace $\mathbf{x}_u$ in the dataset $\mathcal{D}$ with let $\mathbf{x}'_u = \mathbf{x}_u - \boldsymbol{\eta}_u$, where $\boldsymbol{\eta}_u \in \mathbb{R}^d$ is noise sampled IID from the uniform distribution on $[\delta_1, \delta_2]^d$. Denote this updated dataset as $\mathcal{D}'$ and let $\mathcal{X}'$ denote its examples without labels. Then $\mathcal{X}'$ is in general linear position with probability 1, and the global minimum of*

$$L'(\mathbf{w}) = \frac{1}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}} (f_{\mathbf{w}}(\mathbf{x}) - y)^2 \tag{17}$$

is less than or equal to $t\gamma^2/N$ if and only if a set cover of $\mathcal{T}$ exists containing t sets.

The rest of this section is devoted to the proof of theorem E.1. We first prove that $\mathcal{X}'$ is indeed in general position. We then prove both directions of the if and only if statement.

E.3.1 General Position of Perturbed Dataset

Lemma E.2. *The examples of the perturbed dataset $\mathcal{X}'$ are in general linear position.*

Proof. We examine linear rather than affine dependencies between examples since the ReLU that we are using has no bias. Let us partition the training examples as

$$\mathcal{X}' = \mathcal{X}'_U \cup \mathcal{X}_{\mathcal{T}} \cup \{\mathbf{x}_1, \mathbf{x}_\gamma\}. \tag{18}$$

where $\mathcal{X}_{\mathcal{T}}$ corresponds the examples (14) and $\mathcal{X}'_U$ to corresponds to the examples in (15) with perturbations as in theorem E.1.

From their definitions, it is clear that the set of N vectors $\mathcal{X}_{\mathcal{T}} \cup \{\mathbf{x}_1, \mathbf{x}_\gamma\}$ are linearly independent and thus in general position. Since the $\{\boldsymbol{\eta}_u\}_{u\in U}$ are sampled IID from the uniform distribution on $[\delta_1, \delta_2]^d$, the introduction of the examples $\mathcal{X}'_U$ almost surely introduces no new nontrivial linear dependencies. □

E.3.2 Set Cover of Required Size Exists

Now suppose that a set cover $\mathcal{S} \subseteq \mathcal{T}$ of size t exists. Choose parameters $\mathbf{w} \in \mathbb{R}^d$ with coordinates $w_1 = 1$, $w_\gamma = \gamma$, $w_{T_i} = -2$ if $T_i \in \mathcal{S}$, and $w_{T_i} = 0$ otherwise.

Lemma E.3. *The loss on the perturbed dataset is equal to $t\gamma^2/N$ when the parameters are set to $\mathbf{w}$.*

Proof. We start by looking at the examples that are unchanged from the original dataset $\mathcal{D}$ in our perturbed version $\mathcal{D}'$. We see that $f_{\mathbf{w}}(\mathbf{x}_\gamma) = w_\gamma = \gamma = y_\gamma$ and $f_{\mathbf{w}}(\mathbf{x}_1) = w_1 = 1 = y_1$, so these two examples have a loss of zero. When $T_i \in \mathcal{S}$, we have $f_{\mathbf{w}}(\mathbf{x}_{T_i}) = \phi(w_\gamma + w_{T_i}) = \phi(\gamma - 2) = 0$ since $\gamma < 2$, so these examples incur a loss of $y_{T_i}^2/N = \gamma^2/N$. When $T_i \notin \mathcal{S}$, we have $f_{\mathbf{w}}(\mathbf{x}_{T_i}) = \phi(w_\gamma + w_{T_i}) = \phi(\gamma) = \gamma = y_{T_i}$, so these examples have a loss of zero. Overall these examples contribute a total of $t\gamma^2/N$ to the loss.

Now consider the examples $\mathbf{x}'_u \in \mathcal{X}'_U$. By definition, $\mathbf{x}'_u = \mathbf{x}_u - \boldsymbol{\eta}_u = \mathbf{e}_1 + \sum_{T_i \ni u} \mathbf{e}_{T_i} - \boldsymbol{\eta}_u$. The preactivation for such an example is

$$\begin{aligned}\mathbf{w}^T\mathbf{x}'_u &= w_1 + \sum_{T_i \ni u} w_{T_i} - \mathbf{w}^T\boldsymbol{\eta}_u \\ &= 1 - \sum_{u\in T_i\in\mathcal{S}} w_{T_i} 2 - \mathbf{w}^T\boldsymbol{\eta}_u.\end{aligned}$$

Since $\mathcal{S}$ is a set cover of U, we know at least one $T_i \in \mathcal{S}$ exists such that $u \in T_i$. Hence $1 - \sum_{u\in T_i\in\mathcal{S}} w_{T_i} 2 \leq -1$. Recall that the entries of $\boldsymbol{\eta}_u$ are all positive and less than δ_2, and recall that all entries of $\mathbf{w}$ are greater than or equal to -2. Thus, $-\mathbf{w}^T\boldsymbol{\eta}_u \leq 2d\delta_2$. Because we defined δ_2 such that $\delta_2 < \frac{1}{2d}$, we see that $2d\delta < 1$. Hence $-\mathbf{w}^T\boldsymbol{\eta}_u < 1$, so we see that the preactivation is less than $-1 + 1 = 0$. Thus applying a ReLU activation to this preactivation will output a 0. Thus $f_{\mathbf{w}}(\mathbf{x}'_u) = 0 = y'_u$ for all $u \in U$. Hence the examples from $\mathcal{X}'_U$ do not contribute to the loss, and the loss at $\mathbf{w}$ is $t\gamma^2/N$. □

E.3.3 Set Cover of Required Size Does Not Exist

Throughout this section, let $\mathbf{w} \in \mathbb{R}^d$ be parameter values as defined in the previous section. We first prove the following lemma.

Lemma E.4. *If the minimal loss over the original dataset $\mathcal{D}$ is less than or equal to $t\gamma^2/N$, then the minimal loss over the perturbed dataset $\mathcal{D}'$ is less than or equal to $t\gamma^2/N$.*

Proof. From the proof of Theorem 8 in Goel et al. (2020), we know that the minimal loss over the original dataset is less than or equal to $t\gamma^2/N$ only if a set cover of size t exists. This lemma then follows directly from lemma E.3. □

We now have the following.

Lemma E.5. *If the minimal loss over the perturbed dataset $\mathcal{D}'$ is greater than $t\gamma^2/N$, then no set cover of U exists that is comprised of t or fewer sets from $\mathcal{T}$.*

Proof. The contraposition of lemma E.4 states that if the the minimal loss over the perturbed dataset $\mathcal{D}'$ is greater than $t\gamma^2/N$, then the minimal loss over the original dataset $\mathcal{D}$ is greater than $t\gamma^2/N$. From the proof of Theorem 8 in Goel et al. (2020), this implies that no set cover of U exists that consists of t or fewer sets from $\mathcal{T}$. □

F Proof for Upper Bound on Required Overparameterization

In this section, we provide a proof of theorem 4.3. We begin with the following lemma.

Lemma F.1. *Let $A, B \subseteq \mathbb{R}^d$ be finite subsets such that $|B| = d+1$ and the d-th coordinate of any $\mathbf{a} \in A$ is strictly less than the d-th coordinate of any $\mathbf{b} \in B$. Furthermore, assume that their union $A \cup B$ is in general position. For any $\mathbf{b} \in B$, let $y_{\mathbf{b}} \in \mathbb{R}$ be its label. Then there exist parameters $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{d+1}$ satisfying*

$$\mathbf{w}_1^T \bar{\mathbf{a}} \leq 0, \qquad \mathbf{w}_1^T \bar{\mathbf{b}} \geq 0,$$
$$\mathbf{w}_2^T \bar{\mathbf{a}} \leq 0, \qquad \mathbf{w}_2^T \bar{\mathbf{b}} \geq 0,$$

for all $\mathbf{a} \in A$ and $\mathbf{b} \in B$ such that the function

$$f(\mathbf{x}) = \phi(\mathbf{w}_1^T \bar{\mathbf{x}}) - \phi(\mathbf{w}_2^T \bar{\mathbf{x}}) \tag{19}$$

satisfies $f(\mathbf{a}) = 0$ for all $\mathbf{a} \in A$ and $f(\mathbf{b}) = y_{\mathbf{b}}$ for all $\mathbf{b} \in B$.

Proof. Let $\alpha \in \mathbb{R}$ be some value that is strictly greater than the d-th coordinate of any $\mathbf{a} \in A$ and strictly less than the d-th coordinate of $\mathbf{b} \in B$. Such an α is guaranteed to exist based on the assumptions of the lemma. Now let $\mathbf{u} = \mathbf{e}_d - \alpha \mathbf{e}_{d+1} \in \mathbb{R}^{d+1}$. It is clear that $\mathbf{u}^T \bar{\mathbf{a}} < 0$ for all $\mathbf{a} \in A$ and $\mathbf{u}^T \bar{\mathbf{b}} > 0$ for all $\mathbf{b} \in B$. We can thus multiply $\mathbf{u}$ by a positive scalar to get a vector $\tilde{\mathbf{u}} \in \mathbb{R}^{d+1}$ such that $\tilde{\mathbf{u}}^T \bar{\mathbf{a}} < -1$ for all $\mathbf{a} \in A$ and $\tilde{\mathbf{u}}^T \bar{\mathbf{b}} > 1$ for all $\mathbf{b} \in B$.

Let us now look at the unconstrained problem of finding a $\mathbf{w} \in \mathbb{R}^{d+1}$ such that $\mathbf{w}^T \bar{\mathbf{b}} = y_{\mathbf{b}}$ for all $\mathbf{b} \in B$. As B contains $d+1$ examples in general positions, such a $\mathbf{w}$ will always exist and can be found via standard linear regression.

Define

$$\beta_A = \max_{\mathbf{a} \in A} \phi(\mathbf{w}^T \bar{\mathbf{a}}) \tag{20}$$

and

$$\beta_B = \max_{\mathbf{b} \in B} \phi(-\mathbf{w}^T \bar{\mathbf{b}}). \tag{21}$$

Let $\beta = \max\{\beta_A, \beta_B\}$. Then setting $\mathbf{w}_1 = \mathbf{w} + \beta \tilde{\mathbf{u}}$ and $\mathbf{w}_2 = \beta \tilde{\mathbf{u}}$ satisfies the conditions of the lemma. □

Algorithm 3 Modified Greedy Local Search (mGLS) Heuristic

Input: data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, output weights $\mathbf{v} \in \mathbb{R}^{m+1}$, max steps $T \in \mathbb{N}$
$A_0 \in \text{vert}(\mathcal{Z}^m)$ {Random initial zonotope vertex.}
for $t \in \{0, \dots, T\}$ **do**
  $A_{t+1} \leftarrow A_t$
  $W^* \leftarrow$ solution of (3) for A_t
  if any of the constraints in (3) are equalities **then**
    $A_f \leftarrow A_t$ with all those constraints flipped
    $N_f \leftarrow$ subset of neighbors of A_t differing only on one of those constraints
    $\mathcal{N} \leftarrow (\{A_f\}, N_f, \text{neighbors}(A_t) \setminus N_f)$
  else
    $\mathcal{N} \leftarrow (\text{neighbors}(A_t))$
  end if
  for $G \in \mathcal{N}$ **do**
    for $A' \in G$ **do**
      if $\mathcal{L}^*(A'; \mathcal{D}) < \mathcal{L}^*(A_{t+1}; \mathcal{D})$ **then**
        $A_{t+1} \leftarrow A'$
        continue main loop
      end if
    end for
  end for
  if $A_{t+1} = A_t$ **then**
    return A_t
  end if
end for
return A_T

We are now ready for the proof. Let $\mathcal{D}$ and $\mathcal{D}_k$ for $k = 1, \dots, \lceil \frac{N}{d+1} \rceil$ be defined as in section 4.2.1. Let us do a proof by induction on k. Suppose that f is a ReLU network with $2\lceil \frac{N}{d+1} \rceil - 2$ hidden units fitting the labels in $\bigcup_{k'=1}^{k-1} \mathcal{D}_{k'}$ exactly.

Let us now relabel the entire dataset by subtracting the predictions of f from the labels to get

$$\mathcal{D}' = \{(\mathbf{x}, y - f(\mathbf{x})) \mid (\mathbf{x}, y) \in \mathcal{D}\}. \tag{22}$$

Define $\mathcal{D}'_k$ accordingly. Clearly, we have the labels being all zero for all examples in $\mathcal{D}' \setminus \mathcal{D}'_k$. The labels for examples in $\mathcal{D}'_k$ will generally be non-zero.

We can use lemma F.1 to find a unit layer network g such that $g(\mathbf{x}) = 0 = y$ for all $(\mathbf{x}, y) \in \mathcal{D}' \setminus \mathcal{D}'_k$ and $g(\mathbf{x}) = y$ for all $(\mathbf{x}, y) \in \mathcal{D}'_k$. Hence g fits $\mathcal{D}'$ exactly. From this it is clear that $f + g$ fits the original dataset $\mathcal{D}$ exactly. We can find a ReLU network with $2\lceil \frac{N}{d+1} \rceil$ hidden units representing $f + g$ by having its last two units be the units of g and the remaining units be the units of f.

G Modified Greedy Local Search

This section provides more information on the additional heuristics used in the mGLS algorithm introduced in section 4.3.1. The purpose of these modifications is to reduce the typical number of convex problems that we have to solve in a run of the algorithm.

The major difference is that as we iterate over neighboring zonotope vertices, we move to any vertex with a lower loss than the current vertex. This is in contrast to algorithm 2, which evaluates the loss at every neighbor and moves to the one with the lowest loss. Especially near the start of the optimization procedure, we find that this greatly reduces the number of vertices that we need to solve convex programs for. The order in which we iterate over the neighbors is mostly random with the caveat discussed below.

We also make use of geometric information coming the optimal parameter values given the current vertex to preferentially try some subsets of neighboring vertices first. If they lie at the boundary of the current activation region, then it stands to reason that activation regions on the other side

of that boundary are more likely to have better solutions. Solutions lying on the boundary of an activation region have a subset of preactivations that are exactly zero. Equivalently, a subset of the inequalities in (3) become equalities at the solution. In such cases, we first try the vertex that has all of those constraints flipped. Note that this vertex is not usually a neighbor of the current current in the 1-skeleton of the zonotope and might not even be feasible. If feasible and k constraints are flipped, then that vertex and the current vertex belong to the same k-face of the zonotope. We then try the neighbors of the current vertex that correspond to flipping one of those constraints. Afterwards, we try the remaining neighbors.

H Experimental Details

H.1 Synthetic Data

H.1.1 Synthetic Dataset Generation

Here we present the details of the generation of the synthetic datasets used in the experiments in this paper.

We start out with the dimension of the input d and the number of units m_{gen} in the shallow ReLU network used to generate the labels. We use this to calculate the number of examples $N = (d+1)m_{\text{gen}}$. We then generate the examples $\{\mathbf{x}_i\}_{i=1}^N \subseteq \mathbb{R}^d$ by sampling them i.i.d. from a standard normal distribution.

We use a randomly generated ReLU network to label these examples. We can express this network as

$$g(\mathbf{x}) = \mathbf{v}_{\text{gen}}^T \phi(W_{\text{gen}} \bar{\mathbf{x}}) + c_{\text{gen}}. \tag{23}$$

We generate the parameters $\mathbf{v}_{\text{gen}} \in \mathbb{R}^{m_{\text{gen}}}$, $W_{\text{gen}} \in \mathbb{R}^{m_{\text{gen}} \times (d+1)}$, and $c_{\text{gen}} \in \mathbb{R}$ by via sampling from standard normal distributions. The label for the i-th example can then be expressed as $y_i = g(\mathbf{x}_i)$.

H.1.2 Training Details

All experiments on the synthetic datasets used the mean squared error (MSE) loss. The network architecture for these experiments takes the form of

$$f(\mathbf{x}) = \mathbf{v}^T \phi(W \bar{\mathbf{x}}) + c, \tag{24}$$

where $\mathbf{v} \in \mathbb{R}^m$, $W \in \mathbb{R}^{m \times (d+1)}$, and $c \in \mathbb{R}$. We use $m \in \mathbb{N}$ to denote the number of units in the network that we train on the dataset.

Gradient Descent All experiments involving gradient descent on synthetic datasets in this paper used batch gradient descent with a learning rate of 1e-3 for 400,000 steps. All parameters, including the second layer weights $\mathbf{v}, c$, were trained. We used the parameter initialization scheme from Glorot & Bengio (2010).

Random Vertex A random vertex was selected by sampling the first layer weights $W \in \mathbb{R}^{m \times (d+1)}$ from a standard Gaussian and taking its corresponding activation pattern over the dataset. We randomly initialize $\mathbf{v}$ with values chosen uniformly from the set $\{-1, 1\}$. Optimizing over all of the parameters of a shallow ReLU network within a single activation region is non-convex. However, the problems of training W, c given a fixed $\mathbf{v}$ and training $\mathbf{v}, c$ given a fixed W are convex. The former is a slight variant of (3) while the latter is simple linear regression over fixed features. We thus iterate between solving these two problems until we converge to fixed loss value.

This process is guaranteed to converge to a local minima of the loss (Xu & Yin, 2013). However, it is possible that our process of optimizing the second layer weights here is suboptimal and does not reach the global optimum within the activation region.

GLS Heuristic Here we fix $\mathbf{v} \in \mathbb{R}^m$ to have $m/2$ entries set to -1 and $m/2$ entries set to +1. We slightly modify the convex problem (3) to include optimizing the $c \in \mathbb{R}$ in addition to the first layer weights $W \in \mathbb{R}^{m \times (d+1)}$. We choose the starting vertex at random by sampling the first layer weights W from a standard Gaussian and taking its corresponding activation pattern over the dataset. We set $T = 1024$ as the maximum number of steps.

Table A1: Results of all synthetic data experiments performed in this paper. The large numbers are the median final MSE over 16 runs for GLS heuristic and over 8 runs for gradient descent and random vertex. The subscript numbers provide the standard deviation over the runs. Some cells are empty as the particular combination of d, m_{gen}, m, and optimization method was not needed for our set of comparisons.

d	m_{GEN}	m	GRADIENT DESCENT	RANDOM VERTEX	GLS HEURISTIC
4	2	2	$3.82\text{E-}10_{3.2\text{E-}04}$	$1.25\text{E-}01_{1.7\text{E-}01}$	$8.27\text{E-}01_{1.4\text{E}00}$
4	2	3	$3.43\text{E-}11_{7.6\text{E-}10}$	$3.85\text{E-}05_{1.2\text{E-}02}$	–
4	2	4	$6.01\text{E-}12_{3.7\text{E-}09}$	$4.10\text{E-}08_{8.1\text{E-}07}$	$1.58\text{E-}01_{5.3\text{E-}01}$
4	2	8	–	$8.81\text{E-}09_{1.4\text{E-}08}$	$8.29\text{E-}13_{1.7\text{E-}02}$
4	2	16	–	$1.13\text{E-}08_{7.3\text{E-}08}$	$4.11\text{E-}17_{1.2\text{E-}16}$
4	4	4	$8.30\text{E-}03_{1.8\text{E-}01}$	$1.98\text{E-}03_{6.6\text{E-}02}$	$8.11\text{E-}01_{8.5\text{E-}01}$
4	4	6	$5.04\text{E-}12_{5.7\text{E-}10}$	$3.13\text{E-}04_{7.7\text{E-}04}$	–
4	4	8	$6.54\text{E-}12_{9.7\text{E-}10}$	$4.70\text{E-}05_{5.3\text{E-}05}$	$8.47\text{E-}09_{7.8\text{E-}02}$
4	4	16	–	$5.08\text{E-}07_{2.8\text{E-}05}$	$1.68\text{E-}12_{1.6\text{E-}11}$
4	4	32	–	$2.71\text{E-}07_{5.1\text{E-}06}$	$1.97\text{E-}15_{4.7\text{E-}15}$
4	8	8	$8.53\text{E-}03_{5.3\text{E-}03}$	$8.47\text{E-}03_{8.3\text{E-}03}$	$4.68\text{E-}01_{4.0\text{E-}01}$
4	8	12	$2.52\text{E-}11_{4.3\text{E-}10}$	$7.59\text{E-}04_{9.8\text{E-}04}$	–
4	8	16	$2.21\text{E-}11_{2.6\text{E-}09}$	$6.07\text{E-}04_{6.9\text{E-}04}$	$9.96\text{E-}03_{3.1\text{E-}02}$
4	8	32	–	$1.70\text{E-}05_{7.9\text{E-}05}$	$6.68\text{E-}12_{2.9\text{E-}10}$
4	8	64	–	$6.26\text{E-}06_{4.8\text{E-}05}$	$3.49\text{E-}14_{2.8\text{E-}14}$
8	4	4	$2.54\text{E-}03_{1.2\text{E-}01}$	$1.44\text{E-}02_{4.4\text{E-}02}$	$3.74\text{E}00_{2.0\text{E}00}$
8	4	6	$6.11\text{E-}11_{1.4\text{E-}02}$	$1.11\text{E-}03_{2.6\text{E-}02}$	–
8	4	8	$7.69\text{E-}12_{8.2\text{E-}02}$	$6.96\text{E-}08_{4.2\text{E-}05}$	$4.94\text{E-}01_{4.7\text{E-}01}$
8	4	16	–	$2.24\text{E-}08_{2.6\text{E-}08}$	$5.33\text{E-}12_{4.1\text{E-}08}$
8	4	32	–	$6.25\text{E-}09_{2.5\text{E-}09}$	$2.40\text{E-}13_{1.1\text{E-}12}$
8	8	8	$7.47\text{E-}03_{1.2\text{E-}02}$	$2.45\text{E-}02_{9.9\text{E-}03}$	$3.67\text{E}00_{2.7\text{E}00}$
8	8	12	$7.30\text{E-}12_{1.2\text{E-}10}$	$1.69\text{E-}04_{3.1\text{E-}04}$	–
8	8	16	$5.83\text{E-}13_{7.0\text{E-}12}$	$2.07\text{E-}06_{9.5\text{E-}06}$	$1.77\text{E}00_{8.9\text{E-}01}$
8	8	32	–	$8.01\text{E-}08_{1.1\text{E-}07}$	$3.09\text{E-}12_{6.5\text{E-}11}$
8	8	64	–	$2.43\text{E-}08_{6.6\text{E-}09}$	$3.15\text{E-}13_{4.1\text{E-}13}$
8	16	16	–	$2.72\text{E-}02_{3.3\text{E-}02}$	$7.07\text{E}00_{2.7\text{E}00}$
8	16	24	–	$1.61\text{E-}03_{1.8\text{E-}03}$	–
8	16	32	–	$8.57\text{E-}05_{2.1\text{E-}04}$	$1.47\text{E}00_{6.3\text{E-}01}$
8	16	64	–	$5.62\text{E-}07_{4.8\text{E-}07}$	$2.35\text{E-}02_{3.9\text{E-}02}$
8	16	128	–	$1.03\text{E-}07_{7.6\text{E-}08}$	$3.16\text{E-}13_{4.7\text{E-}12}$
16	8	8	–	$2.79\text{E-}02_{4.5\text{E-}02}$	$1.85\text{E}01_{6.4\text{E}00}$
16	8	12	–	$9.05\text{E-}07_{2.2\text{E-}05}$	–
16	8	16	–	$1.10\text{E-}07_{2.0\text{E-}08}$	$8.76\text{E}00_{2.7\text{E}00}$
16	8	32	–	$2.86\text{E-}08_{6.4\text{E-}09}$	$1.74\text{E}00_{8.2\text{E-}01}$
16	8	64	–	$9.53\text{E-}09_{2.2\text{E-}09}$	$2.82\text{E-}13_{7.9\text{E-}11}$
16	16	16	–	$4.99\text{E-}02_{1.6\text{E-}02}$	$2.81\text{E}01_{6.4\text{E}00}$
16	16	24	–	$4.70\text{E-}05_{1.1\text{E-}04}$	–
16	16	32	–	$4.80\text{E-}07_{2.1\text{E-}06}$	$1.63\text{E}01_{3.0\text{E}00}$
16	16	64	–	$8.13\text{E-}08_{1.9\text{E-}08}$	$4.03\text{E}00_{1.0\text{E}00}$
16	16	128	–	$2.75\text{E-}08_{3.0\text{E-}09}$	$3.42\text{E-}13_{3.8\text{E-}12}$

H.1.3 Full Results

We experimented with a range of d, m_{gen}, and m values. Present our full results in table A1. The scores represent the median of 16 runs for random vertex scores and the median of 8 runs for the rest.

H.2 Toy Versions of Real-World Datasets

H.2.1 Dataset Creation

Our datasets were created from the MNIST (LeCun et al., 2010) and Fashion MNIST (Xiao et al., 2017) datasets. Both datasets are 10-way multiclass classification datasets; however, our mGLS algorithm only works for ReLU networks with scalar output. Hence we have to create binary classification tasks from these datasets.

We did this by restricting each dataset to two classes and having the task to correctly differentiate between only those two classes. For MNIST, we chose the 4 and the 9 classes. For Fashion MNIST, we chose the pullover and the coat classes. These classes were chosen for the interclass similarity of their examples, which increases the difficulty of the task.

To reduce the dimensionality of the data, we performed principle components analysis (PCA) using the `scikit-learn` Python package (Pedregosa et al., 2011) on all of the training examples in each dataset belonging to their respective two chosen classes. When then used the first $d \in \{8, 16\}$ whitened components for our dataset. We then took the first $N \in \{350, 700\}$ examples in the training split as our training dataset. We always chose the same examples across experiments to reduce variance.

H.2.2 Training Details

Since these datasets were binary classification tasks, we used the sigmoid cross entropy loss function $\ell(\hat{y}, y) = -y\hat{y} + \sigma(\hat{y})$, where σ is the logistic sigmoid function. The network architecture for these experiments takes the form of

$$f(\mathbf{x}) = \mathbf{v}^T \phi(W\bar{\mathbf{x}}) + c, \tag{25}$$

where $\mathbf{v} \in \mathbb{R}^m$, $W \in \mathbb{R}^{m\times(d+1)}$, and $c \in \mathbb{R}$. We use $m \in \mathbb{N}$ to denote the number of units in the network that we train on the dataset. In all experiments in this section, set $\mathbf{v}$ to a vector containing half ones and half negative ones and froze it throughout training. The rest of the variables were optimized during training. Note that this is different than what we did for the synthetic datasets.

Gradient Descent We trained for one million steps with a learning rate of 1e-3 using batch gradient descent. We used the parameter initialization scheme from Glorot & Bengio (2010).

Random Vertex A random vertex was selected by sampling the first layer weights $W \in \mathbb{R}^{m\times(d+1)}$ from a standard Gaussian and taking its corresponding activation pattern over the dataset. We then solved its corresponding convex program (3) using the ECOS (Domahidi et al., 2013) solver in the `cvxpy` Python package (Diamond & Boyd, 2016). We also optimized the bias $c \in \mathbb{R}$ in the final layer as well as the first layer parameters parameters in the convex program.

mGLS Heuristic We chose the initial vertex by sampling the first layer weights $W \in \mathbb{R}^{m\times(d+1)}$ from a standard Gaussian and taking its corresponding activation pattern over the dataset. Like for the random vertex experiment, we also optimized the second layer bias $c \in \mathbb{R}$ in the convex program. We used the mGLS algorithm presented in appendix G to perform the optimization. We set $T = 2048$ as the maximum number of steps.

H.2.3 Full Results

We experimented with a range of d, N, m values on both MNIST 5/9 and Fashion MNIST coat/pullover. We present our full results comparing gradient descent to the random vertex method in table A2 and comparing gradient descent to mGLS in table A3. Random vertex results are the median of 16 runs while the results for the other two methods are the median of 8 runs.

Table A2: Results of all experiments comparing gradient descent to random vertex optimization in this paper. The subscripts provide standard deviation across runs.

Dataset				Gradient Descent		Random Vertex	
	d	N	m	Loss	Acc (%)	Loss	Acc (%)
MNIST	8	350	4	$1.16\text{E-}01_{1.1\text{E-}02}$	$95.6_{0.80}$	$6.04\text{E-}01_{6.7\text{E-}02}$	$67.4_{8.53}$
	8	350	8	$5.37\text{E-}02_{7.9\text{E-}03}$	$98.9_{0.45}$	$5.27\text{E-}01_{1.1\text{E-}01}$	$72.9_{8.30}$
	8	350	16	$2.09\text{E-}02_{2.6\text{E-}03}$	$99.5_{0.16}$	$3.21\text{E-}01_{7.3\text{E-}02}$	$85.3_{4.87}$
	8	350	32	$8.38\text{E-}03_{5.1\text{E-}04}$	$100.0_{0.00}$	$2.30\text{E-}01_{5.0\text{E-}02}$	$91.4_{2.36}$
	8	350	64	$4.21\text{E-}03_{1.5\text{E-}04}$	$100.0_{0.00}$	$1.71\text{E-}01_{4.9\text{E-}02}$	$94.1_{1.86}$
	8	700	4	$1.62\text{E-}01_{1.9\text{E-}03}$	$93.2_{0.73}$	$6.50\text{E-}01_{8.7\text{E-}02}$	$63.4_{9.13}$
	8	700	8	$1.24\text{E-}01_{6.7\text{E-}03}$	$95.0_{0.56}$	$4.94\text{E-}01_{8.6\text{E-}02}$	$75.9_{7.11}$
	8	700	16	$7.90\text{E-}02_{6.1\text{E-}03}$	$97.6_{0.49}$	$3.95\text{E-}01_{6.4\text{E-}02}$	$82.1_{4.34}$
	8	700	32	$3.97\text{E-}02_{1.5\text{E-}03}$	$99.4_{0.05}$	$2.79\text{E-}01_{5.7\text{E-}02}$	$88.6_{3.02}$
	8	700	64	$1.94\text{E-}02_{6.5\text{E-}04}$	$100.0_{0.04}$	$2.14\text{E-}01_{2.2\text{E-}02}$	$91.5_{1.06}$
	16	350	4	$1.78\text{E-}02_{2.2\text{E-}03}$	$99.5_{0.18}$	$5.95\text{E-}01_{7.8\text{E-}02}$	$68.7_{7.10}$
	16	350	8	$5.75\text{E-}03_{7.4\text{E-}04}$	$100.0_{0.00}$	$4.89\text{E-}01_{1.2\text{E-}01}$	$75.4_{9.28}$
	16	350	16	$2.49\text{E-}03_{2.8\text{E-}04}$	$100.0_{0.00}$	$4.14\text{E-}01_{1.1\text{E-}01}$	$80.0_{7.91}$
	16	350	32	$1.12\text{E-}03_{5.0\text{E-}05}$	$100.0_{0.00}$	$1.39\text{E-}01_{9.1\text{E-}02}$	$95.0_{3.95}$
	16	350	64	$4.87\text{E-}04_{2.2\text{E-}05}$	$100.0_{0.00}$	$4.81\text{E-}02_{3.9\text{E-}02}$	$99.1_{1.29}$
	16	700	4	$3.51\text{E-}02_{4.4\text{E-}03}$	$99.3_{0.18}$	$6.61\text{E-}01_{5.6\text{E-}02}$	$57.7_{7.61}$
	16	700	8	$1.16\text{E-}02_{1.3\text{E-}03}$	$99.9_{0.16}$	$5.34\text{E-}01_{9.0\text{E-}02}$	$72.8_{7.68}$
	16	700	16	$4.61\text{E-}03_{2.9\text{E-}04}$	$100.0_{0.00}$	$4.37\text{E-}01_{8.7\text{E-}02}$	$78.9_{6.01}$
	16	700	32	$2.06\text{E-}03_{1.1\text{E-}04}$	$100.0_{0.00}$	$2.36\text{E-}01_{6.3\text{E-}02}$	$90.1_{3.11}$
	16	700	64	$1.00\text{E-}03_{4.0\text{E-}05}$	$100.0_{0.00}$	$1.20\text{E-}01_{3.3\text{E-}02}$	$95.6_{1.44}$
Fashion MNIST	8	350	4	$2.98\text{E-}01_{1.0\text{E-}02}$	$88.4_{0.67}$	$5.97\text{E-}01_{7.2\text{E-}02}$	$68.4_{8.60}$
	8	350	8	$2.24\text{E-}01_{1.3\text{E-}02}$	$91.4_{0.71}$	$5.58\text{E-}01_{5.5\text{E-}02}$	$72.9_{6.28}$
	8	350	16	$1.39\text{E-}01_{9.8\text{E-}03}$	$95.9_{0.90}$	$4.96\text{E-}01_{5.7\text{E-}02}$	$78.6_{3.39}$
	8	350	32	$6.71\text{E-}02_{7.4\text{E-}03}$	$98.8_{0.43}$	$3.78\text{E-}01_{2.8\text{E-}02}$	$85.1_{1.68}$
	8	350	64	$3.21\text{E-}02_{2.3\text{E-}03}$	$100.0_{0.19}$	$2.91\text{E-}01_{3.1\text{E-}02}$	$88.1_{1.34}$
	8	700	4	$3.52\text{E-}01_{3.8\text{E-}03}$	$84.8_{0.53}$	$6.62\text{E-}01_{5.1\text{E-}02}$	$60.5_{7.12}$
	8	700	8	$3.15\text{E-}01_{6.4\text{E-}03}$	$86.6_{0.63}$	$5.99\text{E-}01_{4.4\text{E-}02}$	$69.3_{4.66}$
	8	700	16	$2.48\text{E-}01_{1.0\text{E-}02}$	$90.5_{0.76}$	$5.17\text{E-}01_{4.2\text{E-}02}$	$76.1_{3.20}$
	8	700	32	$1.65\text{E-}01_{5.0\text{E-}03}$	$94.3_{0.50}$	$4.31\text{E-}01_{3.0\text{E-}02}$	$80.7_{1.78}$
	8	700	64	$8.70\text{E-}02_{6.0\text{E-}03}$	$98.3_{0.37}$	$3.66\text{E-}01_{1.7\text{E-}02}$	$84.4_{1.27}$
	16	350	4	$1.93\text{E-}01_{1.6\text{E-}02}$	$92.4_{1.25}$	$6.24\text{E-}01_{6.2\text{E-}02}$	$64.4_{9.07}$
	16	350	8	$8.33\text{E-}02_{1.3\text{E-}02}$	$98.0_{0.70}$	$5.46\text{E-}01_{6.0\text{E-}02}$	$72.7_{5.67}$
	16	350	16	$3.71\text{E-}02_{3.7\text{E-}03}$	$99.3_{0.14}$	$4.27\text{E-}01_{5.2\text{E-}02}$	$80.1_{3.29}$
	16	350	32	$1.43\text{E-}02_{3.3\text{E-}03}$	$100.0_{0.20}$	$3.00\text{E-}01_{4.4\text{E-}02}$	$87.3_{2.31}$
	16	350	64	$6.72\text{E-}03_{4.3\text{E-}04}$	$100.0_{0.00}$	$1.65\text{E-}01_{1.0\text{E-}01}$	$94.3_{4.22}$
	16	700	4	$2.84\text{E-}01_{1.1\text{E-}02}$	$88.1_{0.48}$	$6.30\text{E-}01_{4.6\text{E-}02}$	$64.8_{5.66}$
	16	700	8	$1.95\text{E-}01_{8.5\text{E-}03}$	$92.0_{0.31}$	$5.95\text{E-}01_{4.9\text{E-}02}$	$68.4_{4.37}$
	16	700	16	$1.13\text{E-}01_{7.5\text{E-}03}$	$96.5_{0.58}$	$5.19\text{E-}01_{4.6\text{E-}02}$	$75.7_{3.27}$
	16	700	32	$4.53\text{E-}02_{3.4\text{E-}03}$	$99.4_{0.25}$	$4.26\text{E-}01_{2.4\text{E-}02}$	$81.4_{1.45}$
	16	700	64	$2.05\text{E-}02_{1.5\text{E-}03}$	$100.0_{0.06}$	$3.43\text{E-}01_{1.5\text{E-}02}$	$85.6_{1.14}$

Table A3: Results of all experiments comparing gradient descent to mGLS in this paper. The subscripts provide standard deviation across runs.

Dataset	m	Gradient Descent		mGLS Heuristic	
		Loss	Acc (%)	Loss	Acc (%)
MNIST	4	$1.16\text{E-}01_{1.1\text{E-}02}$	$95.6_{0.80}$	$1.09\text{E-}01_{3.7\text{E-}02}$	$95.6_{1.76}$
	8	$5.37\text{E-}02_{7.9\text{E-}03}$	$98.9_{0.45}$	$1.95\text{E-}03_{1.5\text{E-}02}$	$100.0_{0.58}$
	16	$2.09\text{E-}02_{2.6\text{E-}03}$	$99.5_{0.16}$	$8.24\text{E-}04_{1.2\text{E-}03}$	$100.0_{0.00}$
	32	$8.38\text{E-}03_{5.1\text{E-}04}$	$100.0_{0.00}$	$3.73\text{E-}02_{8.5\text{E-}03}$	$99.0_{0.61}$
Fashion MNIST	4	$2.98\text{E-}01_{1.0\text{E-}02}$	$88.4_{0.67}$	$2.88\text{E-}01_{1.8\text{E-}02}$	$88.3_{0.64}$
	8	$2.24\text{E-}01_{1.3\text{E-}02}$	$91.4_{0.71}$	$1.72\text{E-}01_{4.5\text{E-}02}$	$93.7_{1.99}$
	16	$1.39\text{E-}01_{9.8\text{E-}03}$	$95.9_{0.90}$	$2.91\text{E-}03_{1.5\text{E-}02}$	$100.0_{0.47}$
	32	$6.71\text{E-}02_{7.4\text{E-}03}$	$98.8_{0.43}$	$2.33\text{E-}02_{6.1\text{E-}02}$	$99.4_{2.45}$