

Title

Optional Subtitle

Andras Filip Plaian

Master's Thesis, Autumn 2021



The study programme is unspecified. Please consult the documentation for the package **masterfrontpage** in order to correctly print the colophon.

Abstract

Add new section about results in

Acknowledgements

Rewrite this.

Contents

Abstract	i
Acknowledgements	iii
Contents	v
List of Figures	vii
1 Introduction	1
2 Compressed Sensing	3
2.1 Main Assumptions: Sparsity and Compressibility	3
2.2 Basic Requirements of A	4
2.3 Finding Suitable Reconstruction Algorithms	5
2.4 Basis Pursuit	9
2.5 Matrix Analysis: Coherence	15
3 Neural Networks	17
3.1 Supervised Machine Learning	17
3.2 Design	17
3.3 Training	19
3.4 The Universal Approximation Theorem	20
4 Instabilities In Deep Learning	21
Appendices	23

List of Figures

3.1	Extremely simplified NN with a single hidden layer.	18
3.2	Different choices of activations functions.	19

CHAPTER 1

Introduction

Write introduction here.

Outline

The rest of the text is organised as follows:

CHAPTER 2

Compressed Sensing

In an underdetermined system of linear equations there are fewer equations than unknowns. In mathematical terms this can be stated as the matrix equation $A\mathbf{x} = \mathbf{y}$ where $A \in \mathbb{C}^{m \times N}$, $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{y} \in \mathbb{C}^m$ and $m < N$. From classical linear algebra it follows that this equation is unsolvable in the general case, however, given certain conditions, it is possible to find an exact or an estimate of a solution.

The research area associated with these assumptions is called Compressed Sensing (CS). The goal of this chapter is to introduce the reader to the needed concepts to be able to determine the needed conditions for being able to find an efficient, stable and robust reconstruction method of an estimate or solution of the *inverse problem* $\mathbf{x} = A^{-1}\mathbf{y}$ by exploiting the theory of CS.

We also want to be able to have some answer to the following questions:

- What properties and conditions does A need to have in order for \mathbf{x} to be reconstructable?
- Does a suitable algorithm for finding \mathbf{x} exist?

2.1 Main Assumptions: Sparsity and Compressibility

The main assumption we do about the vector \mathbf{x} we are trying to recover is that it is sparse. Informally, a vector is *sparse* if most of its components is zero, particularly if we find an appropriate basis for the given vector $\mathbf{y} = A\mathbf{x}$. For instance, natural images are sparse in the wavelet domain. Before we define sparsity formally, we recall the definition of the *support* of a given vector.

Definition 2.1. The support of a vector $\mathbf{x} \in \mathbb{C}^N$ is the index set of its non-zero entries, that is:

$$\text{supp}(\mathbf{x}) := \{i \in [N] : x_i \neq 0\}.$$

In Compressive Sensing it is customary to abuse the l_0 notation to denote the cardinality of the support set, i.e. the number of non-zero entries of a vector. The $\|\cdot\|_0$ fails to be a norm since it does not satisfy the scaling property of norms. We can now define *s-sparse* vectors.

2. Compressed Sensing

Definition 2.2. The vector $\mathbf{x} \in \mathbb{C}^N$ is called *s-sparse* if it has no more than s non-zero entries, that is if: $\|\mathbf{x}\|_0 \leq s$

The notion of sparsity is an ideal one, meaning that in the real world it may very well be that the vector we are trying to recover is only close to being sparse. In order for CS to handle more problems, we may also be interested in the following notion of *compressibility*.

Definition 2.3. For $p > 0$, the measure of a vectors compressibility is given by the l_p -error of best s -term approximation to $\mathbf{x} \in \mathbb{C}^N$ defined by

$$\sigma(\mathbf{x})_p := \inf\{\|\mathbf{x} - \mathbf{z}\|_p, \mathbf{z} \in \mathbb{C}^N \text{ is } s\text{-sparse}\}$$

Informally, a vector is compressible if the l_p -error decays quickly in s . For instance in image processing, if we can describe most of the variance in image data with a few components of a vector.

2.2 Basic Requirements of A

What properties does A need to have in order to have the possibility to reconstruct every s -sparse vector *regardless* of the reconstruction method? For instance, what is the minimal number of rows that A needs to have such that it would be possible to find every s -sparse vector? We have the following result:

Theorem 2.4. For a given matrix $A \in \mathbb{C}^{m \times N}$, the following statements are equivalent:

- a) Every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s -sparse solution of $A\mathbf{z} = A\mathbf{x}$, that is, if $A\mathbf{x} = A\mathbf{z}$ and both \mathbf{z} and \mathbf{x} are s -sparse, then $\mathbf{z} = \mathbf{x}$.
- b) The null space $\ker(A)$ does not contain any $2s$ -sparse vector other than the zero vector, that is, $\ker(A) \cap \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_0 \leq 2s\} = \{\mathbf{0}\}$.
- c) For every $S \subset [N]$ with $\text{card}(S) \leq 2s$, the submatrix A_S is injective.
- d) Every set of $2s$ columns of A is linearly independent.

Proof. b) \implies a) Let \mathbf{x}, \mathbf{z} be s -sparse vectors, not necessarily unique, such that $A\mathbf{x} = A\mathbf{z}$. Then $\mathbf{x} - \mathbf{z}$ is at most $2s$ -sparse and using linearity of A we get $A(\mathbf{x} - \mathbf{z}) = \mathbf{0}$. If b) holds, then the kernel only contains the zero vector of $2s$ -sparse vectors, then the only way $A(\mathbf{x} - \mathbf{z}) = \mathbf{0}$ is true, is if $\mathbf{x} = \mathbf{z}$, thus \mathbf{x} is the unique s -sparse vector.

a) \implies b) Assume that \mathbf{x} is the unique s -sparse vector such that $A\mathbf{x} = A\mathbf{z}$. Let $\mathbf{v} \in \ker(A)$ be $2s$ -sparse. Since \mathbf{v} is $2s$ -sparse, there exist a way to construct \mathbf{v} from $\mathbf{x} - \mathbf{z}$ where \mathbf{x}, \mathbf{z} are both s -sparse and such that they have no components

2.3. Finding Suitable Reconstruction Algorithms

in common. Assuming a) holds, then we have $\mathbf{x} = \mathbf{z} = \mathbf{v} = \mathbf{0}$.

b) \iff c) \iff d) For a $2s$ -sparse vector \mathbf{v} with $S = \text{supp } \mathbf{v}$, if $\mathbf{a}_1, \dots, \mathbf{a}_S$ are the column vectors of the submatrix A_S made from A , we get $A\mathbf{v} = A_S\mathbf{v}_S$. Noting that $S = \text{supp } \mathbf{v}$ ranges through all possible subsets of $[N]$ with $\text{card}(S) \leq 2s$ ranging through all the possible $2s$ -sparse \mathbf{v} vectors. Then in this view, linear independence d), injectivity c) and that the kernel is trivial b) are equivalent statements about A . ■

With a little knowledge about how to construct invertible matrices, we can derive the following result from the above theorem.

Corollary 2.5. *For any integer $N \geq 2s$, there exists a matrix $A \in \mathbb{C}^{m \times N}$ with $m = 2s$ rows such that every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ can be recovered from the vector $\mathbf{y} = A\mathbf{x} \in \mathbb{C}^m$.*

Proof. Let t_1, t_2, \dots, t_N be strictly positive numbers and consider the matrix $A \in \mathbb{C}^{m \times N}$ with $m = 2s$ defined by

$$A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_N \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & \cdots & t_N^{2s-1} \end{bmatrix}$$

Now define a square submatrix $A_S \in \mathbb{C}^{m \times m}$ indexed by $S = \{i_1 < \dots < i_m\}$ as

$$A_S = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_{i_1} & t_{i_2} & \cdots & t_{i_m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{i_1}^{2s-1} & t_{i_2}^{2s-1} & \cdots & t_{i_m}^{2s-1} \end{bmatrix}$$

The given submatrix A_S can be recognized as the transposed *Vandermonde matrix*. A result about Vandermonde matrices is that its determinant, in view of A_S , can be expressed as $\det(A_S) = \prod_{k < l \leq m} (t_{i_l} - t_{i_k})$. Since the t 's were defined to be strictly positive it follows that $\det(A_S) > 0$, and hence A_S is invertible. The Invertible Matrix Theorem then tells us that A_S is injective which means that condition *iii*) in Theorem 2.4 holds and then we have a unique s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ such that $A\mathbf{z} = A\mathbf{x}$. ■

2.3 Finding Suitable Reconstruction Algorithms

Recall that the compressed sensing problem is to find an s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ given $\mathbf{y} = A\mathbf{x}$. Abusing the customary l_0 notation for the number of non-zero entries, we can reformulate this problem as an optimization problem, that is:

$$\text{minimize } \|\mathbf{z}\|_0 \quad \text{subject to } A\mathbf{z} = \mathbf{y}. \quad (P_0)$$

A naive approach to solving this could be to look at it as a combinatorial optimization problem and use brute force to calculate every square system $A_S^* A_S \mathbf{x} = A_S^* \mathbf{y}$, for $\mathbf{x} \in \mathbb{C}^S$ where S runs through all possible subsets of

2. Compressed Sensing

$[N]$ with size s . This might very well work on small sizes of N , but the total amount of subsets the algorithm has to go through is determined by the formula $\binom{N}{s}$. A quick illustration shows that with $N = 1000, s = 10$ we have $\binom{1000}{10} \geq \left(\frac{1000}{10}\right)^{10} = 10^{20}$ linear systems of size 10×10 to solve. Assuming one could solve each system in 10^{-10} seconds, it would still take 10^{10} seconds, ie. several human lifespans, thus the brute force approach is completely unpractical for sufficiently large N .

P_0 is NP-hard

It's not only the brute force approach that is not bounded by a polynomial expression, it might be that there doesn't exist an approach bounded by a polynomial expression at all. In fact, we have the following result:

Theorem 2.6. *The l_0 -minimization problem for an arbitrary matrix $A \in \mathbb{C}^{m \times N}$ and a vector $\mathbf{y} \in \mathbb{C}^m$ is NP-hard.*

Before proving the result, let us recall some of the terminology:

- The class of P problems consists of all decision problems for which there exists a polynomial-time algorithm, ie. an input size bounded by a polynomial expression, for finding a solution.
- The class of NP problems, not to be confused with NP-hard, consists of all decision problems for which there exists a polynomial-time algorithm *certifying* a solution.
- The class of NP-hard problems consists of all problems for which *solving* algorithm could be transformed in polynomial time into a *solving* algorithm for any NP-problem.

This means that if we can show that a given problem solves an other problem belonging to a certain class of problems, then the given problem also belongs to that same class.

Proof. If we can show that a known NP-hard problem can be reduced in polynomial time to the l_0 -minimization problem then l_0 -minimization itself is NP-hard.

Let The Exact Cover by 3-sets problem be our known NP-hard problem. The Exact Cover by 3-sets problem says that given a collection $\{\mathcal{C}_i, i \in [N]\}$ of 3-element subsets of $[m]$, does there exist an exact partition or cover of the set $\{1, 2, \dots, m\} = [m]$?

Let $\{\mathcal{C}_i, i \in [N]\}$ be the collection of 3-element subsets of $[m]$. Define vectors $\mathbf{a}_1, \dots, \mathbf{a}_N \in \mathbb{C}^m$ by

$$a_{ij} = \begin{cases} 1 & \text{if } j \in \mathcal{C}_i, \\ 0 & \text{if } j \notin \mathcal{C}_i. \end{cases}$$

where $j \in [m]$.

Define a matrix $A \in \mathbb{C}^{m \times N}$ and a vector $\mathbf{y} \in \mathbb{C}^m$ by

$$A = [\mathbf{a}_1 \cdots \mathbf{a}_N] \text{ and } \mathbf{y} = [1, \dots, 1]^T.$$

Since $N \leq \binom{m}{3}$, this construction can be done in polynomial time since $\binom{m}{3} = \frac{m(m-1)(m-2)}{3!}$ which is a 3rd degree polynomial in m .

If a vector $\mathbf{z} \in \mathbb{C}^N$ satisfies $A\mathbf{z} = \mathbf{y}$, then all the m components of $A\mathbf{z}$ are nonzero and $\|A\mathbf{z}\|_0 = m$. Since each vector \mathbf{a}_i has exactly 3 nonzero components, the vector $A\mathbf{z} = \sum_{j=1}^N z_j \mathbf{a}_j$ has at most 3 $\|z\|_0$ nonzero components, $\|A\mathbf{z}\|_0 \leq 3\|z\|_0$ and consequently $\|z\|_0 \geq m/3$. Now consider the l_0 -minimization and let $\mathbf{x} \in \mathbb{C}^N$ be the output. We separate two cases:

1. If $\|\mathbf{x}\|_0 = m/3$, then the collection $\{\mathcal{C}_j, j \in \text{supp}(\mathbf{x})\}$ forms an exact cover of $[m]$, for otherwise the m components of $A\mathbf{x} = \sum_{j=1}^N x_j \mathbf{a}_j$ would not all be nonzero.
2. If $\|\mathbf{x}\|_0 > m/3$, then no exact cover $\{\mathcal{C}_j, j \in J\}$ can exist, for otherwise the vector $\mathbf{z} \in \mathbb{C}^N$, defined by $z_j = 1$ if $j \in J$ and $z_j = 0$ if $j \notin J$, would satisfy $A\mathbf{z} = \mathbf{y}$ and $\|\mathbf{z}\|_0 = m/3$, contradicting the l_0 -minimality of \mathbf{x} .

This shows that solving l_0 -minimization problem enables one to solve the Exact Cover by 3-sets problem and consequently l_0 -minimization problem is NP-hard. ■

P_q where $0 < q < 1$ also NP-hard

If P_0 optimization is intractable, would P_q optimization be a better approach? Not according to the following proposition:

Proposition 2.7. l_q -minimization for $0 < q < 1$ is NP-hard.

Proof. The proof goes along the same lines as proving the NP-hardness of l_0 -minimization, that is if l_q -minimization can help to verify that a given NP-hard problem has a solution, that is if the minimum of $\|\mathbf{w}\|_q$ subject to $A\mathbf{w} = \mathbf{y}$, equals n , then l_q -minimization must necessarily also belong to the same class of NP-hard or NP-complete problems.

The *partition problem* consists, given integers a_1, a_2, \dots, a_n , in deciding whether there exists two sets $I, J \subset [n]$ such that $I \cap J = \emptyset$, $I \cup J = [n]$ and $\sum_{i \in I} a_i = \sum_{j \in J} a_j$ for $i, j \in I, J$ respectively.

Define

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n & -a_1 & -a_2 & \cdots & -a_n \\ 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & 0 & \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix} \text{ and } \mathbf{y} = [0, 1, 1, \dots, 1]^T.$$

2. Compressed Sensing

Let \mathbf{x} and \mathbf{z} be the first and second half of the input vector \mathbf{w} to A respectively. That $A\mathbf{w} = \mathbf{y}$ implies that $\sum_{i=1}^n a_i x_i = \sum_{i=1}^n a_i z_i$ and $x_i + z_i = 1$ for all i . We seek to minimize $\sum_{i=1}^n (x_i^p + z_i^p)$ under these constraints.

The minimum value for the objective function when only the constraints $x_i + y_i$ are considered is n , and this occurs if and only if only 0's and 1's are involved. If the minimum $\|\mathbf{w}\|_q$ subject to $A\mathbf{w} = \mathbf{y}$ is n , then it will also be n when only the $x_i + y_i = 1$ constraints are considered. This means that either x_i or z_i is equal to 1, for all i . Let I be the set of i 's where $x_i=1$, and J be the set of i 's where $z_i=1$. We then have $I \cap J = \emptyset$ and $I \cup J = [n]$, ie. a partition on the form we seek.

Defining x_i as 1 when $i \in I$ and defining \mathbf{z} similarly, we obtain a vector \mathbf{w} where all the constraints are fulfilled and where the minimum n is obtained. ■

P_q where $q > 1$

If both l_0 and l_q -minimization for $0 < q < 1$ is NP-hard, how about l_q -minimization for $q > 1$? Unfortunately, this optimization problem fails to recover even 1-sparse vectors. We have the following proposition:

Proposition 2.8. *Let $q > 1$ and let A be a $m \times N$ matrix with $m < N$. Then there exists a 1-sparse vector which is not a minimizer of P_q .*

Proof. Since $m < N$, we have to from linear algebra that the kernel of A is non-trivial. Hence there must exists a $\mathbf{v} \neq \mathbf{0} \in \ker(A)$ such that $A\mathbf{v} = \mathbf{0}$. Assume, for the sake of contradiction, that all standard basis vectors \mathbf{e}_j are minimizers. Choose a j so that $v_j \neq 0$. Then

$$\|\mathbf{e}_j + t\mathbf{v}\|_q^q = |1 + tv_j|^q + \sum_{k \neq j} |tv_k|^q = |1 + tv_j|^q + |t|^q \sum_{k \neq j} |v_k|^q$$

Define two functions such that

$$g_+(t) = (1 + tv_j)^q + t^q \sum_{k \neq j} |v_k|^q$$

$$g_-(t) = (1 + tv_j)^q + (-t)^q \sum_{k \neq j} |v_k|^q$$

For $|t| < 1/v_j$, $\|\mathbf{e}_j + t\mathbf{v}\|_q^q$ coincides with g_+ for $t \geq 0$, and g_- for $t \leq 0$. Taking the derivatives of g_+ and g_- we get

$$g'_+(t) = qv_j(1 + tv_j)^{q-1} + qt^{q-1} \sum_{k \neq j} |v_k|^q$$

$$g'_-(t) = qv_j(1 + tv_j)^{q-1} - q(-t)^{q-1} \sum_{k \neq j} |v_k|^q$$

Taking the limit of the two derivatives and then for $q > 1$ we get

$$\lim_{t \rightarrow 0+} g'_+(t) = \lim_{t \rightarrow 0-} g'_-(t) = qv_j$$

This means that near 0, $\|\mathbf{e}_j + t\mathbf{v}\|_q^q$ has derivative arbitrarily near $qv_j \neq 0$. Since a linear function with derivative qv_j has no minimum near 0, then with $t = 0$, \mathbf{e}_j can't be a minimizer of P_q , which consequently contradicts our assumption that all standard basis vectors \mathbf{e}_j are minimizers of P_q . ■

The special case P_q where $q = 1$

Recall that in an optimization problem we try to minimize or maximize an objective function subject to constraint functions. If all the constraint functions are also convex functions, then the optimization problem becomes a convex optimization problem. Thus both l_0 and l_q -minimization for $0 < q < 1$ are non-convex optimization problems. If all the constraint functions are linear functions, then the following convex optimization problem

$$\text{minimize } \|\mathbf{z}\|_1 \quad \text{subject to } A\mathbf{z} = \mathbf{y}. \quad (P_1)$$

is also a linear program. Efficient algorithms exist for solving LP's. For instance, the Simplex method, behaves like a polynomial-time algorithm for solving real-life LP problems. l_1 -minimization is also known *basis pursuit* and during the next sections we shall see that given certain assumptions on A we can actually solve P_1 and that the recovered solutions will be sparse.

2.4 Basis Pursuit

In order to show how BP can identify the solution to (P_0) , we need to introduce some definitions and theorems, an important property which our matrix must have, is the so-called Null Space Property of a matrix A :

Definition 2.9. A matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the *Null Space Property (NSP)* relative to a set $S \subset \{1, 2, \dots, N\}$ if

$$\min \|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\bar{S}}\|_1 \text{ for all } \mathbf{v} \in \ker A \setminus \{\mathbf{0}\}$$

A is said to satisfy the *Null Space Property of order s* if it satisfies the null space property relative to any set $S \subset \{1, 2, \dots, N\}$ with $|S| \leq s$.

The following theorem shows that the NSP of a matrix is a sufficient condition in order to solve (P_0) .

Theorem 2.10. Given a matrix $A \in \mathbb{C}^{m \times N}$, every vector $\mathbf{x} \in \mathbb{C}^N$ supported on a set S is the unique solution (P_1) with $A\mathbf{x} = \mathbf{y}$ if and only if A satisfies the NSP relative to S .

Furthermore, if the set S varies, then every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique solution to (P_1) with $A\mathbf{x} = \mathbf{y}$ if and only if A satisfies the NSP of order s .

2. Compressed Sensing

Proof. Let S be a fixed index set, and assume that every vector $\mathbf{x} \in \mathbb{C}^N$ supported on this set, is the unique minimizer of (P_1) . From the assumption it follows that for $\mathbf{v} \in \ker A \setminus \{\mathbf{0}\}$, the vector \mathbf{v}_S is the unique minimizer of (P_1) . Since $A(\mathbf{v}_S + \mathbf{v}_{\bar{S}}) = \mathbf{0}$ and $-\mathbf{v}_S \neq \mathbf{v}_{\bar{S}}$, from the minimality assumption we must have that $\|-\mathbf{v}_S\|_1 < \|\mathbf{v}_{\bar{S}}\|_1$. This established the NSP relative to S .

Conversely, assume that NSP relative to S holds. Let $\mathbf{x} \in \mathbb{C}^N$ be supported on S and a vector $\mathbf{z} \in \mathbb{C}^N$, $\mathbf{z} \neq \mathbf{x}$, such that $A\mathbf{z} = A\mathbf{x}$. Following the rules for norms and taking complements for the support of a set, we obtain

$$\|\mathbf{x}\|_1 \leq \|\mathbf{x} - \mathbf{z}_S\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{v}_S\|_1 + \|\mathbf{z}_S\|_1 < \|\mathbf{v}_{\bar{S}}\|_1 + \|\mathbf{z}_S\|_1 = \|-\mathbf{z}_{\bar{S}}\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{z}\|_1.$$

Which shows that \mathbf{x} obtains the unique minimum.

To prove the second part of the theorem, let S vary and assume that every s -sparse vector \mathbf{x} is found by solving (P_1) subject to $A\mathbf{x} = \mathbf{y}$. Let \mathbf{z} be the solution to P_0 subject to $A\mathbf{x} = \mathbf{y}$ then $\|\mathbf{z}\|_0 \leq \|\mathbf{x}\|_0$ so that also \mathbf{z} is s -sparse. But since every s -sparse vector is the unique minimizer of (P_1) , we have that $\mathbf{x} = \mathbf{z}$ and the result follows. ■

Minimum Number Of Rows For Basis Pursuit

From the results above it is clear that if a matrix possesses the NSP of order s , the BP will solve (P_1) . Next we will introduce a theorem that can identify when A has the NSP of order s .

Theorem 2.11. *Given a matrix $A \in \mathbb{C}^{m \times N}$, then every set of $2s$ columns of A is linearly independent if and only if A satisfies the NSP of order s .*

Proof. Assume that every $2s$ columns of A is linearly independent, then from The Invertible Matrix Theorem, we have that the kernel of A does not contain any other $2s$ -sparse vector other than $\mathbf{0}$. Now let \mathbf{x} , and \mathbf{z} be s -sparse with $A\mathbf{z} = A\mathbf{x}$. Then $A(\mathbf{x} - \mathbf{z}) = \mathbf{0}$ and $\mathbf{x} - \mathbf{z}$ is $2s$ -sparse, but since $\ker A \setminus \{\mathbf{0}\}$ is empty, we must have $\mathbf{x} = \mathbf{z}$, but this implies that the NSP of order s holds. Conversely, assume that the kernel of A does not contain any other $2s$ -sparse vector other than $\mathbf{0}$, then for any set S with $\text{card}(S) = 2s$, $A_S \mathbf{x} = \mathbf{0}$ only has the trivial solution and thereby $2s$ linearly independent columns. ■

From these results we can derive that :

Corollary 2.12. *In order for Basis Pursuit to solve (P_0) , the matrix $A \in \mathbb{C}^{m \times N}$ has to satisfy:*

$$m \geq 2s$$

Proof. Assume that it is possible to uniquely recover any s -sparse vector \mathbf{x} from $\mathbf{y} = A\mathbf{x}$. Then, by Theorem 2.4, statement a) holds, and consequently so does d), that is, that every set of $2s$ columns of A is linearly independent. Thus $\text{rank}(A) \geq 2s$, but we also have that the rank of a matrix can not be greater

than the number of rows m , so we must have $\text{rank}(A) \leq m$. Combining and arranging these two inequalities, we get that

$$2s \leq \text{rank}(A) \leq m$$

which implies, $m \geq 2s$, the inequality in the corollary. \blacksquare

Stability

Basis Pursuit features another important property, namely *stability*. This property tackles sparsity defects, that is, to recover a vector $\mathbf{x} \in \mathbb{C}^N$ whose error to the true underlying vector is controlled by its distance to an s -sparse vector.

Definition 2.13. A matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the *stable null space property* with $0 < \rho < 1$ relative to a set $S \subset [N]$ if

$$\|\mathbf{v}_S\|_1 \leq \rho \|\mathbf{v}_{\bar{S}}\|_1 \text{ for all } \mathbf{v} \in \ker A$$

Furthermore, it satisfies the stable null space of order s if it satisfies the stable null space property relative to any set $S \subset [N]$ with $\text{card}(S) \leq s$.

If a matrix satisfies the stable null space property of order s , then the l_1 -error to the true underlying vector \mathbf{x} is given by

$$\|\mathbf{x} - \mathbf{x}'\|_1 \leq \frac{2(1+\rho)}{1-\rho} \sigma_s(\mathbf{x})_1 \quad (2.1)$$

The next result will give us a condition for when, the inequality (2.1) and the stable null space property, will hold.

Theorem 2.14. The matrix $A \in \mathbb{C}^{m \times N}$ satisfies the stable null space property with constant $0 < \rho < 1$ relative to S if and only if

$$\|\mathbf{z} - \mathbf{x}\|_1 \leq \frac{1+\rho}{1-\rho} (\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\|\mathbf{x}_{\bar{S}}\|_1) \quad (2.2)$$

for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ with $A\mathbf{z} = A\mathbf{x}$.

Before proving Theorem 2.14, we show that (2.2) implies (2.1). Let S be the set of the s largest non-negative components of \mathbf{x} , then $\|\mathbf{x}_{\bar{S}}\|_1 = \sigma_s(\mathbf{x})_1$. If \mathbf{x}' is a minimizer of (P_1) , then $\|\mathbf{x}'\|_1 \leq \|\mathbf{x}\|_1$ and $A\mathbf{x}' = A\mathbf{x}$. The right-hand side of the inequality (2.2) with $\mathbf{z} = \mathbf{x}'$ will then be equal to the right hand side of inequality (2.1).

Proof. To prove Theorem 2.14, assume that the matrix A satisfies inequality (2.2) for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ with $A\mathbf{z} = A\mathbf{x}$. Given a vector $\mathbf{v} \in \ker(A)$, since $A\mathbf{v}_{\bar{S}} = A(-\mathbf{v}_S)$, we can apply (2.2) with $\mathbf{x} = -\mathbf{v}_S$ and $\mathbf{z} = \mathbf{v}_{\bar{S}}$. This gives us that

$$\|\mathbf{v}\|_1 \leq \frac{1+\rho}{1-\rho} (\|\mathbf{v}_S\|_1 - \|\mathbf{v}_{\bar{S}}\|_1)$$

2. Compressed Sensing

Rearranging and cancelling equal terms, this can be written as

$$\|\mathbf{v}_S\|_1 \leq \rho \|\mathbf{v}_{\bar{S}}\|_1$$

which we recognize as the stable NSP with constant $0 < \rho < 1$.

Conversely, assume that the matrix A satisfies the stable NSP with constant $0 < \rho < 1$ relative to S . Then for $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ with $A\mathbf{z} = A\mathbf{x}$, we get that $\mathbf{v} = \mathbf{z} - \mathbf{x} \in \ker(A)$ yields

$$\|\mathbf{v}_S\|_1 \leq \rho \|\mathbf{v}_{\bar{S}}\|_1. \quad (2.3)$$

Since $\|\mathbf{v}_{\bar{S}}\|_1 = \|(\mathbf{z} - \mathbf{x})_{\bar{S}}\|_1$, we can rewrite (2.3) into

$$\|\mathbf{v}_{\bar{S}}\|_1 \leq \|\mathbf{z}_{\bar{S}}\|_1 - \|\mathbf{x}_{\bar{S}}\|_1 + \rho \|\mathbf{v}_{\bar{S}}\|_1 + 2\|\mathbf{x}_{\bar{S}}\|_1.$$

Since $\rho < 1$, this can be rewritten as

$$\|\mathbf{v}_{\bar{S}}\|_1 \leq \frac{1}{1-\rho} (\|\mathbf{z}_{\bar{S}}\|_1 - \|\mathbf{x}_{\bar{S}}\|_1 + 2\|\mathbf{x}_{\bar{S}}\|_1).$$

Using (2.3) once again, we chain the inequalities to get,

$$\|\mathbf{v}\|_1 = \|\mathbf{v}_{\bar{S}}\|_1 + \|\mathbf{v}_S\|_1 \leq (1+\rho)\|\mathbf{v}_{\bar{S}}\|_1 \leq \frac{1+\rho}{1-\rho} (\|\mathbf{z}_{\bar{S}}\|_1 - \|\mathbf{x}_{\bar{S}}\|_1 + 2\|\mathbf{x}_{\bar{S}}\|_1)$$

which is (2.2), the desired inequality. ■

Robustness

Basis Pursuit features another important property, namely *robustness*. This property tackles noise.

Definition 2.15. Given $q \geq 1$, the matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the l_q -robust null space property (rNSP), with constants $0 < \rho < 1$ and $\tau > 0$, relative to a set $S \subset [N]$ if

$$\|\mathbf{v}_S\|_q < \frac{\rho}{s^{1-1/q}} \|\mathbf{v}_{\bar{S}}\|_1 + \tau \|A\mathbf{v}\| \text{ for all } \mathbf{v} \in \mathbb{C}^N.$$

Furthermore, it satisfies the l_q -rNSP of order s if it satisfies the l_q -rNSP relative to any set $S \subset [N]$ with $\text{card}(S) \leq s$.

In order to introduce a more general result using l_q -rNSP, we need an intermediate result that holds for the ordinary rNSP, that is setting $q = 1$.

Theorem 2.16. *The matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ relative to S if and only if*

$$\|\mathbf{z} - \mathbf{x}\|_1 \leq \frac{1+\rho}{1-\rho}(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\|\mathbf{x}_{\bar{S}}\|_1) + \frac{2\tau}{1-\rho}\|A(\mathbf{z} - \mathbf{x})\| \quad (2.4)$$

for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$.

Proof. Very similar to the proof of Theorem 2.14, will therefore use results from there when appropriate.

Assume A satisfies (2.4) for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$. Let $\mathbf{v} \in \mathbb{C}^N$ be so that $\mathbf{v} = \mathbf{z} - \mathbf{x}$ with $\mathbf{z} = \mathbf{v}_{\bar{S}}$ and $\mathbf{x} = -\mathbf{v}_S$ which gives

$$\|\mathbf{v}\|_1 \leq \frac{1+\rho}{1-\rho}(\|\mathbf{v}_{\bar{S}}\|_1 - \|\mathbf{v}_S\|_1) + \frac{2\tau}{1-\rho}\|A\mathbf{v}\|.$$

Multiplying by $(1 - \rho)$ on both sides and expanding $\|\mathbf{v}\|_1$ on the left side gives

$$(1 - \rho)(\|\mathbf{v}_S\|_1 + \|\mathbf{v}_{\bar{S}}\|_1) \leq (1 + \rho)(\|\mathbf{v}_{\bar{S}}\|_1 - \|\mathbf{v}_S\|_1) + 2\tau\|A\mathbf{v}\|.$$

Expanding, cancelling equal terms and dividing by 2 on both sides results in

$$\|\mathbf{v}_S\|_1 \leq \rho\|\mathbf{v}_{\bar{S}}\|_1 + \tau\|A\mathbf{v}\|$$

which is the rNSP with constants $0 < \rho < 1$ and $\tau > 0$.

For the converse, assume A satisfies rNSP with constants $0 < \rho < 1$ and $\tau > 0$ relative to S . For $\mathbf{z}, \mathbf{x} \in \mathbb{C}^N$ with $\mathbf{v} = \mathbf{z} - \mathbf{x}$, the rNSP gives

$$\|\mathbf{v}_S\|_1 \leq \rho\|\mathbf{v}_{\bar{S}}\|_1 + \tau\|A\mathbf{v}\|,$$

$$\|\mathbf{v}_{\bar{S}}\|_1 \leq \|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + \|\mathbf{v}_S\|_1 + 2\|\mathbf{x}_{\bar{S}}\|_1$$

Combining these two inequalities and using the rNSP again we arrive at

$$\begin{aligned} \|\mathbf{v}\|_1 &= \|\mathbf{v}_S\|_1 + \|\mathbf{v}_{\bar{S}}\|_1 \leq (1 + \rho)\|\mathbf{v}_S\|_1 + \tau\|A\mathbf{v}\| \\ &\leq \frac{1+\rho}{1-\rho}(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\|\mathbf{x}_{\bar{S}}\|_1) + \frac{2\tau}{1-\rho}\|A\mathbf{v}\|, \end{aligned}$$

which is the desired inequality. ■

Theorem 2.17. *Given $1 \leq p \leq q$, suppose that the matrix $A \in \mathbb{C}^{m \times N}$ satisfies the l_q -robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$,*

$$\|\mathbf{x} - \mathbf{z}\|_p \leq \frac{C}{s^{1-1/p}}(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\sigma_s(\mathbf{x})_1) + Ds^{1/p-1/q}\|A(\mathbf{z} - \mathbf{x})\|,$$

where $C := (1 + \rho)^2/(1 - \rho)$ and $D := (3 + \rho)\tau/(1 - \rho)$.

In order to prove the above result, we need the following little lemma:

2. Compressed Sensing

Lemma 2.18. For any $q > p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$, the inequality

$$\sigma_s(\mathbf{x})_q \leq \frac{c_{p,q}}{s^{1/p-1/q}} \|\mathbf{x}\|_p$$

holds with

$$c_{p,q} = \left[\left(\frac{p}{q} \right)^{p/q} \left(1 - \frac{p}{q} \right)^{1-p/q} \right]^{1/p} \leq 1.$$

Particularly for our proof, with the choice $p = 1$ and $q = p$ gives

$$\sigma_s(\mathbf{x})_p \leq \frac{1}{s^{1-1/p}} \|\mathbf{x}\|_1$$

Proof. Proof is omitted here for sake of readability, but can be found in the appendix. ■

Proof. (Theorem 2.17)

Remark that the l_q -rNSP implies the l_1 -rNSP and the l_p -rNSP ($p \leq q$) in the forms

$$\|\mathbf{v}_S\|_1 \leq \rho \|\mathbf{v}_{\bar{S}}\|_1 + \tau s^{1-1/q} \|A\mathbf{v}\|, \quad (2.5)$$

$$\|\mathbf{v}_S\|_p \leq \frac{\rho}{s^{1-1/q}} \|\mathbf{v}_{\bar{S}}\|_1 + \tau s^{1/p-1/q} \|A\mathbf{v}\|, \quad (2.6)$$

for all $\mathbf{v} \in \mathbb{C}^N$ and all $S \subset [N]$ with $\text{card}(S) \leq s$. In view of inequality (2.5), applying Theorem 2.16 with S chosen as an index set of s largest non-negative components of \mathbf{x} we get

$$\|\mathbf{z} - \mathbf{x}\|_1 \leq \frac{1+\rho}{1-\rho} (\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\sigma_s(\mathbf{x})_1) + \frac{2\tau}{1-\rho} s^{1-1/q} \|A(\mathbf{z} - \mathbf{x})\|. \quad (2.7)$$

Next, we choose S as an index set of s largest non-negative components of $\mathbf{z} - \mathbf{x}$, and apply Lemma 2.18 to it, which yields

$$\|\mathbf{z} - \mathbf{x}\|_p \leq \|(\mathbf{z} - \mathbf{x})_{\bar{S}}\|_p + \|(\mathbf{z} - \mathbf{x})_S\|_p \leq \frac{1}{s^{1-1/p}} \|\mathbf{z} - \mathbf{x}\|_1 + \|(\mathbf{z} - \mathbf{x})_S\|_p.$$

In view of inequality (2.6), we get

$$\begin{aligned} \|\mathbf{z} - \mathbf{x}\|_p &\leq \frac{1}{s^{1-1/p}} \|\mathbf{z} - \mathbf{x}\|_1 + \frac{\rho}{s^{1-1/p}} \|(\mathbf{z} - \mathbf{x})_{\bar{S}}\|_1 + \tau s^{1/p-1/q} \|A(\mathbf{z} - \mathbf{x})\| \\ &\leq \frac{1+\rho}{s^{1-1/p}} \|\mathbf{z} - \mathbf{x}\|_1 + \tau s^{1/p-1/q} \|A(\mathbf{z} - \mathbf{x})\|. \end{aligned} \quad (2.8)$$

Substituting (2.7) into the above inequality (2.8), gives us

$$\begin{aligned} \|\mathbf{z} - \mathbf{x}\|_p &\leq \frac{(1+\rho)^2}{(1-\rho)} \frac{1}{s^{1-1/p}} (\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\sigma_s(\mathbf{x})_1) \\ &\quad + \frac{(3+\rho)}{(1-\rho)} \tau s^{1/p-1/q} \|A(\mathbf{z} - \mathbf{x})\|. \end{aligned}$$

Setting $C := (1 + \rho)^2 / (1 - \rho)$ and $D := (3 + \rho)\tau / (1 - \rho)$ and substituting for it above gives us the desired inequality from Theorem 2.17. ■

2.5 Matrix Analysis: Coherence

It's not easy to verify if a certain matrix is suitable for Basis Pursuit, that is, to check if the given matrix has the NSP. Thus it would be practical if we could find an easy way to calculate a quantity from which we can guarantee the success of the recovery algorithm. The *coherence* is such a quantity.

Definition 2.19. Let $A \in \mathbb{C}^{m \times N}$ be a matrix with l_2 -normalized columns, $\mathbf{a}_1, \dots, \mathbf{a}_N$ such that $\|\mathbf{a}_i\|_2 = 1$ for all $i \in [N]$. The *coherence* $\mu = \mu(A)$ of the matrix A is defined as

$$\mu := \max_{1 \leq i \neq j \leq N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$$

The l_1 -coherence μ_1 is defined for $s \in [N - 1]$ as

$$\mu_1(s) := \max_{i \in N} \max \left\{ \sum_{j \in S} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|, S \subset [N], \text{card}(S) = s, i \notin S \right\}$$

From the definition of coherence we have the following result:

Theorem 2.20. Let $A \in \mathbb{C}^{m \times N}$ be a matrix with l_2 -normalized columns. If

$$\mu_1(s) + \mu_1(s - 1) < 1,$$

then every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the vector $\mathbf{y} = A\mathbf{x}$ via Basis Pursuit.

Proof. If we can show that for a given matrix $A \in \mathbb{C}^{m \times N}$, with l_2 -normalized columns satisfying $\mu_1(s) + \mu_1(s - 1) < 1$ also possess the NSP of order s , then according to Theorem 2.10, every s -sparse vector \mathbf{x} recovered via Basis Pursuit, is unique.

Let $\mathbf{a}_1, \dots, \mathbf{a}_N$ denote the columns of A . If $\mathbf{v} \in \ker(A)$, then $A\mathbf{v} = \mathbf{0}$ can be rewritten as the sum of the column vectors of A as the sum $\sum_{j=1}^N v_j \mathbf{a}_j = \mathbf{0}$. Since the $\mu_1(s)$ coherence function is a statement about the sum of the inner products of pairwise columns of A , it will be convenient to rewrite a particular weight v_i as:

$$v_i = v_i \langle \mathbf{a}_i, \mathbf{a}_i \rangle = - \sum_{j=1, j \neq i}^N v_j \langle \mathbf{a}_j, \mathbf{a}_i \rangle$$

Splitting $[N]$ into two parts S and \bar{S} and taking the absolute value of v_i it follows that

$$|v_i| \leq \sum_{l \in \bar{S}} |v_l| |\langle \mathbf{a}_l, \mathbf{a}_i \rangle| + \sum_{j \in S, j \neq i} |v_j| |\langle \mathbf{a}_j, \mathbf{a}_i \rangle|$$

2. Compressed Sensing

Summing over all $i \in S$ we get

$$\begin{aligned}\|\mathbf{v}_S\|_1 &= \sum_{i \in S} |v_i| \leq \sum_{l \in \bar{S}} |v_l| \sum_{i \in S} |\langle \mathbf{a}_l, \mathbf{a}_i \rangle| + \sum_{j \in S} |v_j| \sum_{i \in S, i \neq j} |\langle \mathbf{a}_j, \mathbf{a}_i \rangle| \\ &\leq \sum_{l \in \bar{S}} |v_l| \mu_1(s) + \sum_{j \in S} |v_j| \mu_1(s-1) = \mu_1(s) \|\mathbf{v}_{\bar{S}}\|_1 + \mu_1(s-1) \|\mathbf{v}_S\|_1\end{aligned}$$

Rearranging and neglecting the terms between the first term and the last in the equations above we get

$$(1 - \mu_1(s-1)) \|\mathbf{v}_S\|_1 \leq \mu_1(s) \|\mathbf{v}_{\bar{S}}\|_1$$

using the first part of the condition in the theorem, $\mu_1(s) + \mu_1(s-1) < 1$, we get the desired inequality

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\bar{S}}\|_1$$

which concludes the proof. ■

CHAPTER 3

Neural Networks

Neural Networks (NN's), a biologically-inspired programming paradigm, provide among the best solutions to many problems where the neural network has a lot of training data. Consequently, NN's perform well in areas such as image classification, speech recognition, natural language processing, and so forth. However, NN's have reportedly seen to suffer from instabilities in image reconstruction. Our goal is to get mathematical insight to why these instabilities occur. We begin by introducing the basic elements of NN's in framework of supervised machine learning.

3.1 Supervised Machine Learning

In supervised machine learning, the objective is to learn a mapping f from an input space \mathbb{U} to an output space \mathbb{V} . The 2-tuple set $\{(u_i, v_i)\}_{i=1}^n$, $u_i \in \mathbb{U}$, $v_i \in \mathbb{V}$ is referred to as the *training set*. By varying the output space \mathbb{V} , we can choose what our function f should model.

For instance, if v_i takes the values -1 or 1, then f is modelling a *binary classification* problem or if v_i takes three or more discrete values then f is modelling the *multiclass classification* problem.

Image reconstruction can be cast as a supervised learning problem as follows: Let $\mathbb{U} \subseteq \mathbb{C}^M$ and $\mathbb{V} \subseteq \mathbb{C}^N$ such that $\{(y_i, x_i)\}_{i=1}^n$ is the training set with x_i as images and y_i the corresponding measurements. Our goal is then to find a matrix A such that we can express

$$y_i = Ax_i + \epsilon, \quad i \in \{1, \dots, n\}$$

which corresponds to learning the mapping $f(u_i) = v_i + \epsilon$, $i \in \{1, \dots, n\}$. It is desirable that this mapping has *small training error*, i.e. $f(u_i) \approx v_i$ for $i = 1, \dots, n$, but also that the mapping *generalizes* well to new data $\{(y_k, x_k)\}$ which is close to, but not part of the training set.

The task of computing the mapping f is known as *training the model*, such a mapping could for instance be a NN.

3.2 Design

The term neural network covers a large class of models and learning methods. Here we try to give a description that is sufficient for the later analysis.

3. Neural Networks

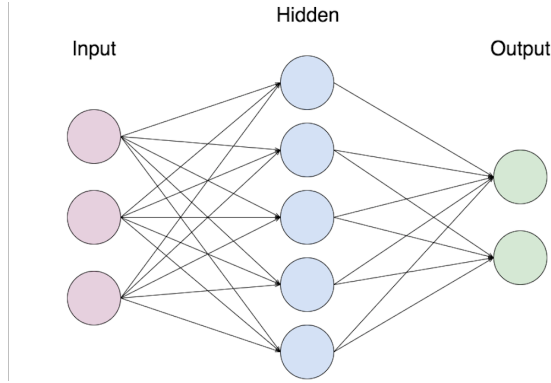


Figure 3.1: Extremely simplified NN with a single hidden layer.

Definition 3.1. Let $n_0, \dots, n_L \in \mathbb{N}$. Let $W_l : \mathbb{C}^{n_{l-1}} \rightarrow \mathbb{C}^{n_l}$ for $l = 1, 2, \dots, L$ be affine maps. Let $\rho_1, \dots, \rho_L : \mathbb{C} \rightarrow \mathbb{C}$ be non-linear, differentiable functions and let $n_0 = M$ and $n_L = N$. Then a map $\phi : \mathbb{C}^M \rightarrow \mathbb{C}^N$ given by

$$\phi(y) = \rho_L(W_L(\dots \rho_1(W_1(y)) \dots))$$

is called a *Neural Network*.

The type of network defined above is a feedforward network. It takes an input y and *feeds* it forward through the network via an alternating sequence of affine maps and non-linear activation functions. It is common to visualize the network as a graph, where the nodes of the graph are the *neurons* and the entries of W_l are the weights assigned to each edge of the graph. At a given *layer*, each neuron takes in a sum of linear combination from the outputs of the neurons in the previous layer, applies a bias, applies an activation function and finally feeds it forward to the next layer of neurons. The number of layers and the number of neurons per layer together with the the activation function ρ is called the *architecture* of the network.

The choice of activation function may be important to the performance of the network. Common activation functions in the literature include:

$$\begin{aligned} ReLU(x) &= \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \\ \tanh(x) &= \frac{e^{2x} - 1}{e^{2x} + 1} \\ \sigma(x) &= \frac{1}{1 + e^{-x}} \end{aligned}$$

If we look closer at the sigmoid function and the hyperbolic tangent, we observe that for large negative input or large positive input, the derivative of these functions will be close to zero. This is known as the *vanishing gradient problem*, we shall refer to it in the next section.

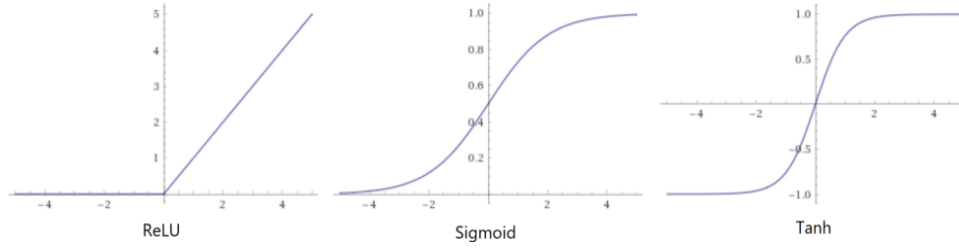


Figure 3.2: Different choices of activations functions.

3.3 Training

The first step in training a NN is to find weights and biases such that the network performs well on the training set. A way to measure how well the network is performing is by defining a *cost function*.

Loss function

Several cost functions may be defined, a popular choice is the *quadratic* cost function, also known as *mean squared error (MSE)*:

$$C(w) = \frac{1}{2n} \sum_{i=1}^n \|\phi(u_i) - v_i\|^2 \quad (3.1)$$

Here, w denotes the collection of all weights and biases in the network, n is the total number of training inputs, v_i is the vector of outputs from the network when u_i is input, and the sum is over all training inputs, u_i . Next, we want to optimize w and b , that is to solve:

$$\min C(w) \quad (3.2)$$

This problem is nonlinear, nonconvex and there is generally no possibility of computing global minimizers of it. Even worse, since a NN needs a lot of parameters to perform well and a lot of training data, it is a major computational challenge. However, there is a practical way to reduce the computational cost, known as *Stochastic Gradient Descent*.

Stochastic Gradient Descent

To make NN's training practical regarding the computational demand, a first-order optimization method is used to solve (ref). These are based on gradient descent. In a gradient descent method we pick a starting point w_0 , then we follow the path of steepest descent, that is, we produce the sequence:

$$w_{i+1} = w_i - \tau_i \nabla C(w_i), \quad i = 0, 1, \dots \quad (3.3)$$

Calculating the gradient requires one pass through all the training set. When the amount of training data is large, this ceases to be computationally feasible. One alternative is already mentioned, stochastic gradient descent. There are

3. Neural Networks

several variations, but the main idea is to compute the gradient from a smaller subset of the training data and thereby reducing the computation needed at step i . In order to apply the stochastic gradient descent method, we need to compute the gradient $C(w_i)$ including all the weights in the network. The common algorithm for computing such gradients, is known as *backpropagation*.

3.4 The Universal Approximation Theorem

In supervised machine learning we are assuming that there exists an underlying function that we are approximating. The Universal Approximation Theorem gives justification for the use of NN's to approximate such a function. Before we state the theorem, we recall the definition of a *discriminatory* function.

Definition 3.2. Let $M(S)$ be the space of finite, signed Borel measures on the compact set S . We say ρ is *discriminatory* if for $\mu \in M(S)$ and

$$\int_{I_n} \rho(\mathbf{w}^T \mathbf{x} + b) d\mu(x) = 0$$

for all $\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}$ then $\mu = 0$.

We are now ready to present the Universal Approximation Theorem and its proof.

Theorem 3.3. (*Universal Approximation Theorem*) Let $C(S)$ be the space of continuous functions on S .

If the activation function ρ in the neural network definition is a continuous, discriminatory function, then the set of all neural networks \mathcal{N} is dense in $C(S)$.

Proof. To prove that \mathcal{N} is dense in $C(S)$, we will prove that its closure is $C(S)$. By way of contradiction, suppose that $\overline{\mathcal{N}} \neq C(S)$, then $\overline{\mathcal{N}}$ is a closed, proper subspace of $C(S)$. By the Hahn-Banach Theorem, there exist a bounded linear functional L , such that $L(\mathcal{N}) = L(\overline{\mathcal{N}}) = 0$ but $L \neq 0$.

By the Riesz Representation Theorem, we can write the functional L as

$$L(h) = \int_S h(x) d\mu(x)$$

for some $\mu \in M(S)$ and for all $h \in C(S)$. In particular, since by definition any NN is a member of \mathcal{N} , and L is identically zero on \mathcal{N} , we have

$$\int_S h(x) d\mu(x) = 0$$

Since we ρ is discriminatory, we have $\mu = 0$. This contradicts that $L \neq 0$.

Therefore, \mathcal{N} is dense in $C(S)$. ■

CHAPTER 4

Instabilities In Deep Learning

4.1 Lack Of Kernel Awareness

Suppose we have trained a NN to approximate an inverse map, $\Psi : \mathbb{C}^m \rightarrow \mathbb{C}^N$ where $m \ll N$. For the NN to be useful, it should perform well on the training set and test set. However, if the NN recovers two vectors \mathbf{x}, \mathbf{x}' whose difference lies in the null space of A , then there exist a ball around $\mathbf{y} = A\mathbf{x}$ where the NN may perform badly with small perturbations from \mathbf{y} . Before we state this consequence as a theorem, recall that local ϵ -Lipschitz constant of a function ϕ at $\mathbf{y} \in \mathbb{C}^m$ is defined as

$$L^\epsilon(\phi, y) = \sup_{0 < d_2(z, y) \leq \epsilon} \frac{d_1(\phi(z), \phi(y))}{d_2(z, y)}$$

Theorem 4.1. (*Universal Instability Theorem*) Let d_1 and d_2 be metrics on \mathbb{C}^N and \mathbb{C}^m respectively, $A : \mathbb{C}^N \rightarrow \mathbb{C}^m$ a linear map, and $\Psi : \mathbb{C}^m \rightarrow \mathbb{C}^N$ a continuous reconstruction map. Suppose there exist $x, x' \in \mathbb{C}^N$ and $\eta > 0$ such that

$$d_1(\Psi(Ax), x) < \eta, \quad d_1(\Psi(Ax'), x') < \eta, \quad (4.1)$$

$$d_2(Ax, Ax') \leq \eta. \quad (4.2)$$

Then there is a closed non-empty ball $\mathcal{B} \subset \mathbb{C}^m$ centred at $y = Ax$ such that the local ϵ -Lipschitz constant at any $\tilde{y} \in \mathcal{B}$ is bounded from below:

$$L^\epsilon(\Psi, \tilde{y}) \geq \frac{1}{\epsilon}(d_1(x, x') - 2\eta), \quad \epsilon \geq \eta. \quad (4.3)$$

Proof. By the definition of the supremum we have that

$$L^\epsilon(\Psi, y) = \sup_{0 < d_2(z, y) \leq \epsilon} \frac{d_1(\Psi(z), \Psi(y))}{d_2(z, y)} \geq \frac{d_1(\Psi(Ax), \Psi(Ax'))}{d_2(Ax, Ax')}$$

Applying the reverse triangle inequality twice we get

$$\frac{d_1(\Psi(Ax), \Psi(Ax'))}{d_2(Ax, Ax')} \geq \frac{d_1(x, \Psi(Ax')) - d_1(\Psi(Ax), x)}{d_2(Ax, Ax')} \geq \frac{d_1(x, x') - d_1(\Psi(Ax), x) - d_1(\Psi(Ax'), x')}{d_2(Ax, Ax')}$$

Applying assumption (4.1) and making the bound sharper with $\epsilon \geq \eta$, we get

4. Instabilities In Deep Learning

$$\frac{d_1(x, x') - d_1(\Psi(Ax), x) - d_1(\Psi(Ax'), x')}{d_2(Ax, Ax')} > \frac{d_1(x, x') - 2\eta}{\epsilon}$$

Neglecting the middleterms we get

$$L^\epsilon(\Psi, y) > \frac{d_1(x, x') - 2\eta}{\epsilon}.$$

Next, let $\eta_1 = d_1(\Psi(Ax), x)$, and observe that $\eta_1 < \eta$ by (4.1).

Since Ψ is continuous at $y = Ax$, there exist a $\delta > 0$ such that for all $\tilde{y} \in \mathbb{C}^m$ with $d_2(y, \tilde{y}) < \delta$, we have that $d_1(\Psi(y), \Psi(\tilde{y})) \leq \eta - \eta_1$. Thus, for a specific $\delta > 0$, we get that $d_1(\Psi(y), \Psi(\tilde{y})) \leq \eta - \eta_1$ for all $\tilde{y} \in \mathcal{B}_1$, where $\mathcal{B}_1 = \{\tilde{y} \in \mathbb{C}^m : d_2(y, \tilde{y}) < \delta\}$. Next, let $\mathcal{B}_2 = \{\tilde{y} \in \mathbb{C}^m : d_2(Ax', \tilde{y}) \leq \eta\}$ and $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2$. Observe that $\mathcal{B} \neq \emptyset$, since $y \in \mathcal{B}$. Thus, for any $\tilde{y} \in \mathcal{B}$ we have

$$\begin{aligned} L^\epsilon(\Psi, \tilde{y}) &= \sup_{0 < d_2(z, \tilde{y}) \leq \epsilon} \frac{d_1(\Psi(z), \Psi(\tilde{y}))}{d_2(z, \tilde{y})} \geq \frac{d_1(\Psi(Ax'), \Psi(\tilde{y}))}{d_2(Ax', \tilde{y})} \\ &\geq \frac{d_1(\Psi(Ax), \Psi(Ax')) - d_1(\Psi(Ax), \Psi(\tilde{y}))}{d_2(Ax', \tilde{y})} \\ &\geq \frac{d_1(x, x') - d_1(\Psi(Ax), x) - d_1(\Psi(Ax'), x') - d_1(\Psi(Ax), \Psi(\tilde{y}))}{d_2(Ax', Ax)} \\ &\geq \frac{d_1(x, x') - \eta_1 - \eta - (\eta - \eta_1)}{d_2(Ax', Ax)} \\ &\geq \frac{1}{\epsilon}(d_1(x, x') - 2\eta). \end{aligned}$$

which is the desired inequality (4.3). ■

Appendices
