

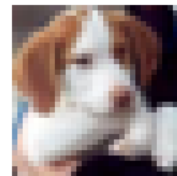
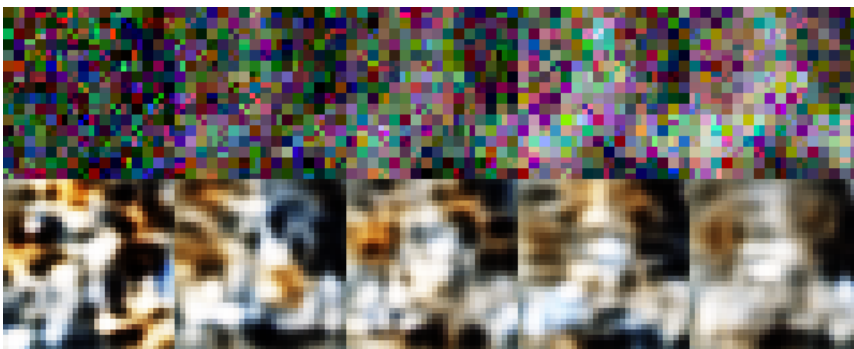


Utrecht University

# One-bit Compressed Sensing with Generative Models

Master's thesis - Mathematical Sciences

Jasper M. Everink



**Supervisor:**  
Dr. Sjoerd Dirksen

**Second reader:**  
Dr. Palina Salanevich

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Size and complexity of signal sets</b>	<b>9</b>
<b>3</b>	<b>(Sub-)Gaussian setting</b>	<b>19</b>
<b>4</b>	<b>Linear and affine covering numbers of signal sets</b>	<b>26</b>
<b>5</b>	<b>Sub-exponential setting</b>	<b>32</b>
<b>6</b>	<b>Heavier tailed setting</b>	<b>47</b>
<b>7</b>	<b>Numerical experiments</b>	<b>57</b>
<b>8</b>	<b>Conclusion and further research</b>	<b>75</b>
<b>A</b>	<b>Miscellaneous proofs</b>	<b>77</b>
<b>B</b>	<b>High-dimensional probability theory</b>	<b>83</b>
	<b>Bibliography</b>	<b>87</b>

# Chapter 1

## Introduction

The reconstruction of a signal  $t_0$  in  $T \subseteq \mathbb{R}^n$  from possibly noisy linear measurements  $y_i$  of the form

$$y_i = \langle a_i, t_0 \rangle + \xi_i, \quad \text{for } i = 1, \dots, m,$$

or, equivalently,

$$y = At_0 + \xi,$$

with measurement vectors  $a_i \in \mathbb{R}^n$ , measurement matrix  $A \in \mathbb{R}^{m \times n}$  and noise  $\xi \in \mathbb{R}^m$ , is a very common problem in applied mathematics, for example, in regression analysis, inverse problems and signal processing. While it is common to work with overdetermined systems, having less measurements can be useful, if not necessary, when data is expensive to obtain. However, underdetermined systems can lead to imprecise solutions if no additional constraints are considered, for example, when  $T = \mathbb{R}^n$ .

In the field of compressed sensing, the standard structural assumption is that the original signal is (nearly-)sparse. That is, in some basis the signal  $t_0$  is in (or near) the set of  $s$ -sparse vectors in  $\mathbb{R}^n$ , i.e.,

$$t_0 \in \Sigma_s^n := \{t \in \mathbb{R}^n : |\{i : t_i \neq 0\}| \leq s\}.$$

This sparsity assumption has been shown to greatly reduce the number of measurements required to reconstruct a signal since various works by Candes and Tao [7, 6]. Although it can be difficult to construct explicit examples of measurement matrices  $A$  with the desired property to require a small number of measurements, probabilistic arguments in the form of random matrices have been very successful. For example, if the measurement matrix  $A$  has i.i.d. sub-Gaussian elements, then there are with high probability good reconstruction guarantees if  $m \geq Cs \ln(en/s)$  for some universal constant  $C$  that only depends on the sub-Gaussian parameter [14].

A lot of progress has been made on the use of random matrices in compressed sensing and its variants. This includes the use of distributions with tails heavier than sub-Gaussian and the consideration of matrices that are more structured than having independently sampled elements or rows [14].

## 1.1 One-bit compressed sensing

An important variant is quantized compressed sensing [8], where the measurements are discretized by a quantizer  $Q : \mathbb{R} \rightarrow \mathcal{Q}$  for some countable set  $\mathcal{Q}$ , resulting in measurements of the form

$$y_i = Q(\langle a_i, t_0 \rangle + \xi_i), \quad \text{for } i = 1, \dots, m.$$

Such quantization is a necessary procedure for computers to be able to process the measurements. When the quantization is done finely, for example, conversion to a 32-bit floating point format, then this problem can still be analyzed using existing techniques from compressed sensing by considering the effect of the quantization as noise, i.e.,

$$y_i = \langle a_i, t_0 \rangle + \xi_i + \eta_i,$$

where

$$\eta_i := Q(\langle a_i, t_0 \rangle + \xi_i) - \langle a_i, t_0 \rangle - \xi_i.$$

This "noise"  $\eta_i$  will be small for a fine quantizer, but not when we take the quantization to the extreme. The extreme case considered in this thesis is one-bit compressed sensing, first introduced by Boufounos et al. [4], where the quantizer is the sign function, resulting in measurements of the form

$$y_i = \text{sign}(\langle a_i, t_0 \rangle + \xi_i), \quad \text{for } i = 1, \dots, m. \quad (1.1)$$

Due to the measurements  $y_i$  being reduced to values in  $\{-1, 1\}$ , a lot of information is lost. Without the one-bit quantization, one would have knowledge of the measurement vector  $a_i$  and the (possible noisy) signed distance to the hyperplane with  $a_i$  as normal. The one-bit quantization changes the signed distance to only the sign, hence only giving us knowledge of  $a_i$  and information on which side of the hyperplane the signal resides.

If we have full knowledge of the noise, then the binary measurement vector  $y$  describes which cell of a tessellation of the signal set by various hyperplanes the signal lies in. See Figure 1.1 for an illustration. Finding the signal can then only be done up to the cell it is contained in, thus good measurement vectors and noise should result in cells of the tessellation that have small diameter. This relates the problem of one-bit compressed sensing to hyperplane tessellation. However, we generally do not have full knowledge of the noise, yet we will see later that partially controlling the noise can be beneficial for the reconstruction.

In the noiseless case, for every  $\lambda > 0$  we have,

$$y_i = \text{sign}(\langle a_i, t_0 \rangle) = \text{sign}(\langle a_i, \lambda t_0 \rangle),$$

hence the magnitude  $\|t_0\|_2$  cannot be reconstructed from noiseless, one-bit measurements. It is still possible to reconstruct the signal up to magnitude from the one-bit measurements by maximizing the correlation between quantized measurements

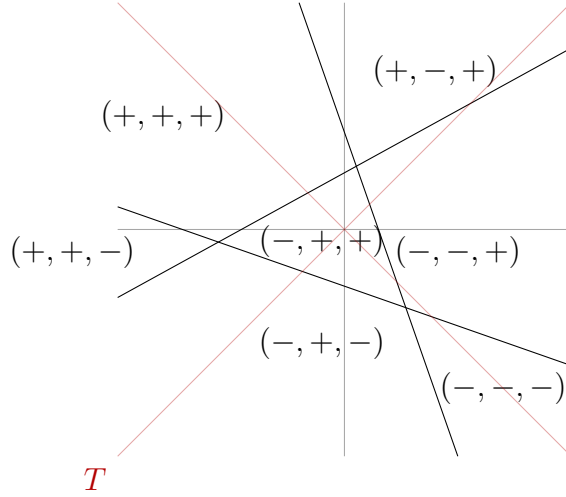


Figure 1.1: Three hyperplanes split a signal set  $T$  into smaller parts, which shows the relation between one-bit compressed sensing and hyperplane tessellation. For simplicity, +1 And  $-1$  are abbreviated to  $+$  and  $-$  respectively.

of the true signal and unquantized measurements, instead of minimizing an error of the form  $\|Az - y\|$ , all possibly with regularization. More precisely, let  $T \subseteq \mathbb{R}^n$  and  $t_0 \in T$  be the true signal, then we can consider the following optimization problem:

$$\max_{t \in T} \langle y, At \rangle, \quad (1.2)$$

or equivalently

$$\max_{t \in T} \sum_{i=1}^m y_i \langle a_i, t \rangle.$$

For such an optimization problem we have the following recovery guarantee.

**Theorem 1.1.1** (Corollary 1.2 from [29]). *Let  $T \subseteq B_2^n$  and fix  $t_0 \in T \cap S_2^{n-1}$ . Let  $a_i$  be i.i.d standard normal random vectors in  $\mathbb{R}^n$ . Denote by  $\hat{t}$  the optimizer of optimization problem (1.2) with the  $y_i$  as defined in equation (1.1). If*

$$m \geq C\epsilon^{-2}w(T)^2,$$

for some  $\epsilon > 0$ , then with probability at least  $1 - 8 \exp(-c\epsilon^2 m)$ ,

$$\|t_0 - \hat{t}\|_2^2 \leq \frac{\epsilon}{\lambda},$$

where  $C > 0$  is a universal constant,  $\lambda$  depends on the noise and  $w(T)$  is called the Gaussian width which is defined by

$$w(T) := \mathbb{E} \sup_{t \in T} \langle g, t \rangle,$$

where  $g$  is a standard Gaussian random vector in  $\mathbb{R}^n$ .

In the case where  $T = \Sigma_s^n$ , the lower bound on the number of measurements is satisfied when

$$m \geq C\epsilon^{-2}s \log(en/s),$$

which is similar to the sub-Gaussian result for unquantized compressed sensing mentioned earlier. Various similar results, including both pre-and post-quantization noise, that are similar to the theorem above can be found in a work by Plan and Vershynin [29] and a version of Lemma 1.1.1 which will be discussed in Chapter 3.

The restriction of recovering only normalized signals is not a problem in certain applications, but, maybe surprisingly, "noise" can actually help reconstruct the magnitude. A signal that lies close to the origin needs little noise for the measurement  $y_i$  to flip sign, while a signal that lies far from the origin needs lots of noise to cause a sign flip. Therefore, it can be practical to add noise before quantization in order to recover the magnitude of the signal. For this setting, consider measurements of the form

$$y_i = \text{sign}(\langle a_i, t_0 \rangle + \xi_i + \tau_i), \quad \text{for } i = 1, \dots, m, \quad (1.3)$$

where  $\xi_i$  is natural noise and  $\tau_i$  is artificial noise, specially chosen to help with reconstructing the norm of  $t_0$ . A naive way of using this artificial noise is to take multiple measurements with the same  $a_i$  and varying  $\tau_i$  to deduce the distance from the origin, therefore attempting to recover the distance lost by one-bit quantization. This method is quite unpractical and we will see later in this thesis that we do not even need to know the values  $\tau_i$  if we can choose the distribution.

Adding artificial pre-quantization noise like above is referred to as dithering and has been successfully used in various works [20, 12, 30] to get guarantees on the reconstruction of both the direction and the magnitude of the signal.

## 1.2 Compressed sensing with generative models

Besides the use of quantization, other structural assumptions on the signal set  $T$  have received a lot of attention recently. Instead of assuming that the signal (in some basis) lies in  $\Sigma_s^n$ , we can also assume that the signal lies in the range of a function whose domain is low-dimensional. Specifically, let  $k \ll n$  and  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$ . We will refer to this function  $G$  as a generative model and assume that the signal lies in its range  $G(X)$ . In later chapters we will show that the number of measurement required for good reconstruction depends on how large  $k$  is and the complexity of  $G$  in the form of the Lipschitz constant of  $G$  and/or the radius of the range. Such a structural assumption was first proposed for unquantized compressed sensing by Bora et al. [2] where a recovery guarantee similar to the following was shown.

**Theorem 1.2.1** (Slight generalization of the main theorem in [2]). *Let  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  be Lipschitz continuous with Lipschitz constant  $\gamma$ ,  $\epsilon > 0$  and  $A$  be a random  $\mathbb{R}^{m \times n}$  matrix with i.i.d. normal entries with mean zero and variance  $1/m$ .*

Take  $x_0 \in X$  fixed and consider measurements of the form  $y = AG(x_0) + \eta$  with fixed noise  $\eta$ . Let  $\hat{x}$  be the minimizer of the optimization problem

$$\min_{x \in X} \|AG(x) - y\|_2^2.$$

If

$$m \geq C \left( \log \mathcal{N} \left( X, \frac{\epsilon}{\gamma} \right) + k \right),$$

then, with probability at least  $1 - 2e^{-cm}$ ,

$$\|G(\hat{x}) - G(x_0)\|_2 \leq 8\|\eta\|_2 + 4\epsilon,$$

where  $c, C > 0$  are universal constants and  $\mathcal{N}(T, \delta)$  is the covering number of  $T$ , defined as the smallest number of balls with radius  $\delta$  required to cover  $T$ .

Two popular choices for generative models are generative adversarial networks and the decoder part of an autoencoder. In this thesis we will only consider the latter for the numerical experiments at the end. An autoencoder is an artificial neural network consisting of an encoder followed by a decoder [15]. By letting the intermediate encoding be relatively low dimensional, the information goes through a bottleneck, hence when the network is trained to approximate the identity function, the signal gets compressed. An autoencoder is illustrated in Figure 1.2. If successfully trained on a data set, then the range of the decoder can be used as a low-dimensional approximation for that data set.

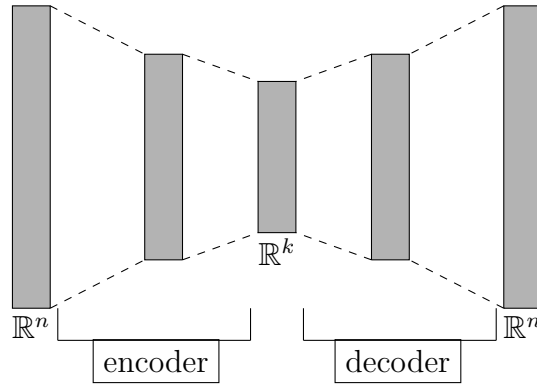


Figure 1.2: Illustration of an autoencoder. The encoder compresses the signal and the decoder decompresses the signal.

Often a data set can be considered as a low dimensional manifold, therefore we can consider the range of a generative model to be an approximation of this manifold. Training a neural network on a data set would then correspond to learning this manifold. One can also work directly with this manifold, for which various recovery guarantees combined with one-bit measurements exist [17, 9].

## 1.3 This work

As the name of the thesis suggests, this thesis will focus on the combination of one-bit measurements with the structural assumption of a generative model. Thus, we have measurements of the form

$$y_i = \text{sign}(\langle a_i, G(x_0) \rangle + \xi_i + \tau_i), \quad \text{for } i = 1, \dots, m,$$

for some generative model  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$ , natural noise  $\xi$  and artificial noise  $\tau$ .

A basic understanding of high-dimensional probability theory is indispensable for understanding this work and the book *High-Dimensional Probability: An Introduction with Applications in Data-Science* by Roman Vershynin [32] is used as a basis for this thesis. A small summary of important definitions and theorems from high-dimensional probability can be found in Appendix B. Furthermore, some machine learning theory will be used in Chapter 4.

In the two theorems mentioned in this introduction, Theorem 1.2.1 and 1.1.1, various quantities measuring the size and complexity of the signal set are used. In Chapter 2 we will look into these quantities in more detail, specifically when the signal set is the range of a generative model that is Lipschitz continuous. Understanding these quantities will allow us to derive recovery guarantees when the measurements and noise are (sub-)Gaussian from results on general signal sets. This is part of Chapter 3 together with a discussion on the (near-)optimality of these guarantees.

In Chapter 4 we will discuss another quantity describing the complexity of a signal set. We will require this quantity for the discussion of a generalization of a result and corresponding proof by Qiu et al. [30] in Chapter 5, which allows the measurement vectors and noise to be sub-exponential. Chapter 6 consists of generalizing the allowed noise to distributions with tails far heavier than sub-exponential and discusses an attempt to further generalize the measurement vectors to certain heavy tailed distributions.

Finally, Chapter 7 consists of various numerical experiments on several aspects of the recovery of a signal from one-bit quantized measurements, both with and without dithering, and comparing the use of a generative model to the sparsity assumption. Most of these experiments will make use of the MNIST data set [25] of handwritten digits, because its simplicity allows us to easily work with various generative models. We will finish with showing that the methods discussed will also work for slightly more complex data sets like CIFAR-10 [22], which consists of small colour images.



## Chapter 2

### Size and complexity of signal sets

In the introduction we have seen that the required number of measurements, also known as the sampling complexity, depends on various quantities that measure the size and complexity of the signal set  $T$ . In Theorem 1.2.1 this is the covering number  $\mathcal{N}(T, \epsilon)$  and in Theorem 1.1.1 this is the Gaussian width  $w(T)$ . Other related quantities will be introduced later in this chapter and Chapter 4. For the derivation of various results on one-bit compressed sensing with generative models in the next chapter we would like to understand these quantities when the signal set is the range of a generative model, i.e.,  $T = G(X)$ . This is the goal of this chapter.

The class of generative models generally considered in the literature is the set of Lipschitz continuous functions, i.e., functions  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  such that, for any  $x, y \in X$ ,

$$\|G(x) - G(y)\|_2 \leq \gamma \|x - y\|_2,$$

for some  $\gamma \in (0, \infty)$ , or equivalently

$$\text{diam}(G(Z)) \leq \gamma \text{diam}(Z),$$

for any  $Z \subseteq X$ . A Lipschitz continuous function with constant  $\gamma$  will also be called a  **$\gamma$ -Lipschitz generative model** in this thesis. This is the same class of functions as considered by Bora et al. [2] and in Theorem 1.2.1. However, for the results by Qiu et al. [30], which will be discussed in more detail in Chapter 4 and 5, a different class of generative models will be considered.

In the next few sections we will see that the Lipschitz continuous functions are a natural choice for generative models, as they behave well with respect to the covering number and Gaussian width. Furthermore, most generative models of interest are Lipschitz continuous, for example, neural networks with Lipschitz continuous activation functions.

**Lemma 2.0.1** (Lemma 8.5 in [2]). *Let  $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$  be an  $l$ -layer feedforward neural network with  $M$ -Lipschitz activation functions and at most  $n$  nodes per layer, i.e.,*

$$G(x) = \sigma_l(b_l + A_l \sigma_{l-1}(\dots \sigma_1(b_1 + A_1 x))),$$

*where the  $\sigma_i$  are all  $M$ -Lipschitz and the matrices  $A_i$  are of size at most  $n \times n$ . For such a generative model, the Lipschitz constant is at most  $(Mna_{\max})^l$ , where  $a_{\max}$  is the largest absolute coefficient in the matrices  $A_i$ .*

## 2.1 Covering Number and Metric Entropy

A set  $\mathcal{N} \subseteq T$  such that any point in  $T$  is at most distance  $\epsilon > 0$  from  $\mathcal{N}$  is called an  $\epsilon$ -**net** of  $T$ . The smallest cardinality of such an  $\epsilon$ -net is called the **covering number**  $\mathcal{N}(T, \epsilon)$  and an  $\epsilon$ -net with such cardinality is called a minimal  $\epsilon$ -net. The covering number and  $\epsilon$ -net are essential concepts within the study of high-dimensional probability, see for example the book by Vershynin [32], and will be used throughout this thesis.

The covering number often arises in arguments in which concentration inequalities are combined with union bounds to obtain bounds for all the points in a net and then approximately extended to the whole set. In such an argument it is not uncommon to find the logarithm of the covering number, i.e.,  $\log_2 \mathcal{N}(T, \epsilon)$ . This quantity is called the **metric entropy** because it represents the number of bits needed to encode all the points in a minimal  $\epsilon$ -net. We will see that it more often arises in cases where it is preferred to work with exponentials or from maximum inequalities like

$$\mathbb{E} \max_{i \in [N]} |X_i| \leq C \max_{i \in [N]} \|X_i\|_{\psi_2} \sqrt{\log N}, \quad (2.1)$$

where  $C > 0$  is a universal constant.

Now consider the range of a  $\gamma$ -Lipschitz generative model  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$ . Due to  $G$  being  $\gamma$ -Lipschitz, any  $\epsilon$ -net  $\mathcal{N}$  of  $X$  can be transformed to a  $\gamma\epsilon$ -net  $G(\mathcal{N})$  of  $G(X)$ . From this observation we directly get the following lemma.

**Lemma 2.1.1.** *If  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  is  $\gamma$ -Lipschitz, then, for any  $\epsilon > 0$ ,*

$$\mathcal{N}(G(X), \epsilon) \leq \mathcal{N}\left(X, \frac{\epsilon}{\gamma}\right).$$

A common choice for the latent space  $X$  is the unit ball  $B_2^k$  for which the following volumetric bound holds.

**Lemma 2.1.2** (Section 4.2 of [32]). *For any  $R > 0$  and  $\epsilon > 0$ ,*

$$\left(\frac{R}{\epsilon}\right)^k \leq \mathcal{N}(RB_2^k, \epsilon) \leq \left(\frac{2R}{\epsilon} + 1\right)^k.$$

Furthermore, for any subset  $T \subseteq RB_2^k$ ,

$$\mathcal{N}(T, \epsilon) \leq \mathcal{N}(RB_2^k, \epsilon/2) \leq \left(\frac{4R}{\epsilon} + 1\right)^k.$$

Combining Lemmas 2.1.1 and 2.1.2 above results in the following standard bound for the covering number of the range of a Lipschitz continuous generative model.

**Corollary 2.1.3.** *Let  $G : B_2^k \rightarrow \mathbb{R}^n$  be  $\gamma$ -Lipschitz. Then, for any  $\epsilon > 0$ ,*

$$\mathcal{N}(G(B_2^k), \epsilon) \leq \left(\frac{2\gamma}{\epsilon} + 1\right)^k.$$

## 2.2 Gaussian Width

Recall that we define the **Gaussian width** of a set  $T \subseteq \mathbb{R}^n$  as

$$w(T) := \mathbb{E} \sup_{t \in T} \langle g, t \rangle,$$

where  $g$  is a standard  $n$ -dimensional Gaussian random vector. From this definition it might not be directly clear why it is called the Gaussian width as it might look more like a "Gaussian radius". However, due to  $g$  being origin symmetric, we can relate this definition to another frequently used definition for the Gaussian width, namely

$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2} \mathbb{E} \sup_{t, s \in T} \langle g, t - s \rangle.$$

The right most term more clearly represents Gaussian width as illustrated in Figure 2.1.

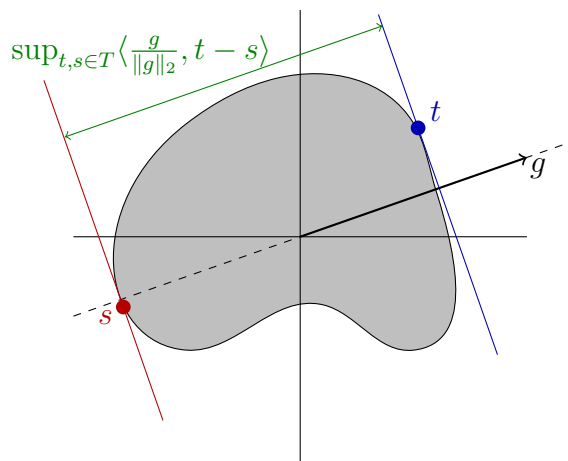


Figure 2.1: Visualization of Gaussian width.

In this thesis, the Gaussian width will appear either as a result of generic chaining or concentration arguments, see for example Lemma 2.2.1 below or Lemma 3.1.2.

**Lemma 2.2.1** (Generic chaining, Corollary 8.6.3 from [32]). *Let  $T \subseteq \mathbb{R}^n$  and  $(X_t)_{t \in T}$  be a mean zero random process. If for all  $t, s \in T$*

$$\|X_t - X_s\|_{\psi_2} \leq K \|t - s\|_2,$$

*for some constant  $K > 0$ , then*

$$\mathbb{E} \sup_{t \in T} X_t \leq CK w(T),$$

*where  $C > 0$  is a universal constant.*

The Gaussian width behaves similarly to the diameter under the transformation by a Lipschitz continuous function. In fact, one can characterize Lipschitz continuity this way.

**Lemma 2.2.2.** *Let  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$ , then  $G$  is  $\gamma$ -Lipschitz if and only if*

$$w(G(Z)) \leq \gamma w(Z), \quad \text{for all } Z \subseteq X.$$

The key to the proof of Lemma 2.2.2 is the Sudakov-Fernique inequality for comparing two Gaussian processes.

**Theorem 2.2.3.** *(Theorem 7.2.11 from [32]) Let  $(X_t)_{t \in T}$  and  $(Y_t)_{t \in T}$  be two mean zero Gaussian processes. If for all  $t, s \in T$ ,*

$$\mathbb{E}(X_t - X_s)^2 \leq \mathbb{E}(Y_t - Y_s)^2,$$

then

$$\mathbb{E} \sup_{t \in T} X_t \leq \mathbb{E} \sup_{t \in T} Y_t.$$

*Proof of Lemma 2.2.2.* Assume that  $G$  is  $\gamma$ -Lipschitz, take any  $Z \subseteq X$  and define the Gaussian processes  $X_t := \langle G(t), g \rangle$  and  $Y_t := \gamma \langle t, g' \rangle$  for all  $t \in Z$ , where  $g \sim N(0, I_n)$  and  $g' \sim N(0, I_k)$ .

Note that these Gaussian processes have mean zero. Furthermore, the second condition of the Sudakov-Fernique inequality is obtained through the observation that

$$\mathbb{E}(\langle G(t) - G(s), g \rangle)^2 = \|G(t) - G(s)\|_2^2 \leq \gamma^2 \|t - s\|_2^2 = \mathbb{E}(\gamma \langle t - s, g' \rangle)^2,$$

hence

$$\mathbb{E}(X_t - X_s)^2 \leq \mathbb{E}(Y_t - Y_s)^2.$$

Thus by the Sudakov-Fernique inequality we can conclude that

$$w(G(Z)) = \mathbb{E} \sup_{t \in Z} \langle G(t), g \rangle = \mathbb{E} \sup_{t \in Z} X_t \leq \mathbb{E} \sup_{t \in Z} Y_t = \gamma \mathbb{E} \sup_{t \in Z} \langle t, g' \rangle = \gamma w(Z).$$

For the other direction, let  $T = \{t, s\} \subset X$  be any two-point subset of  $X$ , using the assumption that  $w(G(T)) \leq \gamma w(T)$  combined with  $w(T) = \frac{1}{2} w(T - T)$  we find

$$\mathbb{E} \max\{0, g_{\|G(t) - G(s)\|_2}\} \leq \gamma \mathbb{E} \max\{0, g_{\|t - s\|_2}\},$$

where  $g_\sigma$  are centered Gaussian random variables with variance  $\sigma^2$ . Rewriting both sides to

$$\|G(t) - G(s)\|_2 \mathbb{E} \max\{0, N(0, 1)\} \leq \gamma \|t - s\|_2 \mathbb{E} \max\{0, N(0, 1)\},$$

and dividing both sides by  $\mathbb{E} \max\{0, N(0, 1)\} \neq 0$  completes the proof. ■

Just like for the covering number, we still need to know the Gaussian width of the standard latent space, the unit ball  $B_2^k$ . This can be derived as follows:

$$w(B_2^k) = \mathbb{E} \sup_{t \in B_2^k} \langle g, t \rangle = \mathbb{E} \|g\|_2 \leq \sqrt{\mathbb{E} \|g\|_2^2} = \sqrt{k}.$$

Thus we can conclude the following bound for the Gaussian width of the range of a Lipschitz continuous generative model.

**Corollary 2.2.4.** *If  $G : B_2^k \rightarrow \mathbb{R}^n$  is  $\gamma$ -Lipschitz, then*

$$w(G(B_2^k)) \leq \gamma \sqrt{k}.$$

While these results are the best we can achieve only assuming the Lipschitz continuity of the generative model, better results can be achieved with the additional assumption that the radius of the range of the generative model is relatively small. This is often the case when working with images, where the generative model does not stray too far from the origin, yet the Lipschitz constant can be very large. For example, the images in the MNIST data set consist of  $28^2$  pixels with values in  $[0, 1]$  and therefore the data set lies in a ball of radius 28, hence we expect that a generative model trained for this data set does not stray far outside this  $l_2$ -ball. These improved bounds will be discussed at the end of the next section, as the proof follows directly from bounds for a variant of the Gaussian width which we will discuss first.

## 2.3 Localized Gaussian Width

A slight variant on the Gaussian width is the **localized Gaussian width**, which, for a set  $T \subseteq \mathbb{R}^n$  and distance  $\rho > 0$ , is defined as

$$w((T - T) \cap \rho B_2^n).$$

Intuitively, it measures how well signals that are within distance  $\rho$  of each other are separated by standard Gaussian measurements, but the main reason of its importance will be discussed in the next section.

Combining the monotonicity of Gaussian width with bounds derived in the previous section, we get for  $\gamma$ -Lipschitz generative models  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  the simple bound

$$w((G(X) - G(X)) \cap \rho B_2^n) \leq \min\{2\gamma w(X), \rho\sqrt{n}\}.$$

However, this relatively straightforward bound ruins the advantage of the relatively low dimensional size of  $G(X)$  together with the small size of  $\rho B_2^n$ . So the rest of this section will be used to derive a better bound that uses both of these properties. To do this, we will use the covering number discussed earlier. We begin by answering the question how to extend the Gaussian width of a net to the whole signal set. For this, we will use the following approximation property of the Gaussian width.

**Lemma 2.3.1.** *Let  $T, S \subset \mathbb{R}^n$  such that  $S \subseteq T + \epsilon B_2^n$ , i.e., every point in  $S$  is at most distance  $\epsilon$  from  $T$ , then*

$$w(S) \leq w(T) + \epsilon\sqrt{n}.$$

*Proof.* Denote by  $T(s)$  the closest point in  $T$  to  $s$ . By assumption we have that  $\|T(s) - s\|_2 \leq \epsilon$  for all  $s \in S$ . We then get that

$$\begin{aligned} w(S) &= \mathbb{E} \sup_{s \in S} \langle g, s \rangle \\ &\leq \mathbb{E} \sup_{s \in S} \langle g, T(s) \rangle + \mathbb{E} \sup_{s \in S} \langle g, s - T(s) \rangle \\ &\leq w(T) + w((T - S) \cap \epsilon B_2^n) \\ &\leq w(T) + \epsilon w(B_2^n) \\ &\leq w(T) + \epsilon\sqrt{n}, \end{aligned}$$

concluding the proof. ■

A simple corollary of this lemma is the continuity of Gaussian width. If  $S, T \subset \mathbb{R}^n$  are equivalent up to  $\epsilon$ -thickening, i.e.,  $S \subseteq T + \epsilon B_2^n$  and  $T \subseteq S + \epsilon B_2^n$ , then Lemma 2.3.1 implies

$$|w(S) - w(T)| \leq \epsilon\sqrt{n},$$

thus similar sets have similar Gaussian width.

Now we can use Lemma 2.3.1 to prove the following improved bound on the localized Gaussian width.

**Lemma 2.3.2.** *Let  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  be  $\gamma$ -Lipschitz, then*

$$w((G(X) - G(X)) \cap \rho B_2^n) \leq C\rho \left( \sqrt{\log \left( \mathcal{N} \left( X, \frac{\rho}{2\sqrt{n}\gamma} \right) \right)} + 1 \right),$$

for some universal constant  $C > 1$ .

*Proof.* Let  $\mathcal{N}$  be a minimal  $\epsilon/2\gamma$ -net of  $X$ , then  $G(\mathcal{N})$  is an  $\epsilon/2$ -net of  $G(X)$  and  $G(\mathcal{N}) - G(\mathcal{N})$  is an  $\epsilon$ -net of  $G(X) - G(X)$ . Furthermore, we have that  $\log |G(\mathcal{N}) - G(\mathcal{N})| \leq 2 \log |\mathcal{N}|$ . Using Lemma 2.3.1 we find

$$w((G(X) - G(X)) \cap \rho B_2^n) \leq w((G(\mathcal{N}) - G(\mathcal{N})) \cap (\rho + \epsilon) B_2^n) + \epsilon\sqrt{n}.$$

Note the additional  $\epsilon$  in the distance of the localized Gaussian width. By using a bound on the Gaussian width of finite sets (see for example Proposition 7.29 in [14] or inequality (2.1)), we find

$$\begin{aligned} w((G(X) - G(X)) \cap \rho B_2^n) &\leq C(\rho + \epsilon) \sqrt{\log |G(\mathcal{N}) - G(\mathcal{N})|} + \epsilon\sqrt{n} \\ &\leq C(\rho + \epsilon) \sqrt{2 \log |\mathcal{N}|} + \epsilon\sqrt{n} \\ &= C(\rho + \epsilon) \sqrt{2 \log \mathcal{N}(X, \epsilon/2\gamma)} + \epsilon\sqrt{n}. \end{aligned}$$

Choosing  $\epsilon = \rho/\sqrt{n}$  allows us to conclude that

$$w((G(X) - G(X)) \cap \rho B_2^n) \leq C\rho \left( \sqrt{\log \left( \mathcal{N} \left( X, \frac{\rho}{2\sqrt{n}\gamma} \right) \right)} + 1 \right),$$

for some  $C > 1$ . ■

To get back to the final comment in the previous section, let us assume that  $G$  is a bounded Lipschitz generative model, then using a similar argument as in the proof of Lemma 2.3.2 above we get the following result.

**Lemma 2.3.3.** *Let  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  be  $\gamma$ -Lipschitz and assume that  $G(X) \subseteq RB_2^n$  for some  $R > 0$ , then*

$$w(G(X)) = w(G(X) \cap RB_2^n) \leq CR \left( \sqrt{\log \left( \mathcal{N} \left( X, \frac{R}{\sqrt{n}\gamma} \right) \right)} + 1 \right),$$

for some universal constant  $C > 1$ .

For the case  $X = B_2^k$ , the bound results in

$$w(G(B_2^k)) \leq CR \left( \sqrt{k \log \left( \frac{2\sqrt{n}\gamma}{R} + 1 \right)} + 1 \right).$$

This bound is more efficient than the bound from Corollary 2.2.4 when the generative model has a large Lipschitz constant, but the radius of the range is small. While it might be possible to reduce the logarithmic dependency on  $n$ , practically,  $\gamma$  will be much larger than  $n$  and will therefore be the dominating term in the logarithm.

The dependency on  $\gamma$  cannot be fully removed. It is straightforward to construct a sequence of Lipschitz continuous functions such that the convex hulls of their range converges to  $B_2^n$ , hence, by continuity of Gaussian width and because the Gaussian width is invariant under taking the convex hull, the low dimensionality of the Gaussian width would be lost while the Lipschitz constant becomes much larger.

For comparison, when  $T$  is a compact Riemannian manifold similar results exist, depending logarithmically on the diameter, volume and reach of the manifold.

**Theorem 2.3.4** (Theorem 3.3. from [17]). *Let  $\mathcal{M} \subset \mathbb{R}^n$  be a compact  $k$ -dimensional Riemannian manifold, then*

$$w(\mathcal{M}) \leq C \operatorname{diam}(\mathcal{M}) \sqrt{k \max \left\{ 1, \log \left( c \frac{\sqrt{k}}{\min\{1, \operatorname{reach}(\mathcal{M})\}} \right) \right\}} + \log(\max\{1, \operatorname{vol}(\mathcal{M})\}),$$

for some constants  $c, C > 0$ .

## 2.4 Star-shaped Sets

The need for the localized Gaussian width comes from the following lemma used for a result by Dirksen et al. [12] which will be discussed in the next chapter.

**Lemma 2.4.1.** *Let  $f : \mathbb{R}^n \rightarrow [0, \infty)$  be positive homogeneous,  $\rho > 0$  and  $W \subseteq \mathbb{R}^n$  satisfy  $\lambda w \in W$  for all  $w \in W$  and  $\lambda \in [0, 1]$ , then*

$$\sup_{w \in W: \|w\|_2 \geq \rho} f(w/\|w\|_2^2) = \sup_{w \in W: \|w\|_2 = \rho} f(w)/\rho^2 \leq \sup_{w \in W: \|w\|_2 \leq \rho} f(w)/\rho^2.$$

This lemma is used for  $W = T - T$ , thus bounding the supremum over signals that are far away by a supremum over the signals that are close together. The final supremum can later be bounded by localized Gaussian width.

A problem lies in the additional requirement on  $T - T$ , hence the focus of this section is to analyse this requirement, specifically in the Lipschitz continuous generative model case where  $T = G(X)$ .

The requirement on  $W$  is referred to as star-shapedness and is defined as follows.

**Definition 2.4.2** (Star-shaped). A set  $T \subset \mathbb{R}^n$  is called **star-shaped around the origin** when

$$\lambda T \subseteq T, \quad \text{for all } \lambda \in [0, 1],$$

and called **star-shaped around**  $t \in \mathbb{R}^n$  if  $T - t$  is star-shaped around the origin. Similarly to convex sets, we define  $\text{Star}_0(T)$  and  $\text{Star}_t(T)$  to be the smallest star-shaped set around the origin and around the point  $t \in \mathbb{R}^n$  that contains  $T$ , respectively.

Because (localized) Gaussian width is monotone, it will be enough to bound  $G(X) - G(X)$  by its origin star-shaped hull, hence the quantity of interest in this section will be

$$w(\text{Star}_0(G(X) - G(X)) \cap \rho B_2^n).$$

In certain special cases this star-shaped hull is not necessary. This can happen when  $G(X)$  is already star-shaped around some point, because then  $G(X) - G(X)$  is star-shaped around the origin. For example, when  $G$  is positive homogeneous and  $X$  is star-shaped around the origin, then  $G(X)$  is also star-shaped around the origin. A practical example is when  $G$  is an unbiased feedforward neural networks with ReLU activation with  $X = B_2^k$ , as considered by Qiu et al. [30].

If  $G(X) - G(X)$  is not already star-shaped around the origin, then a way to upper bound it is to use that  $\text{Star}_0(T - T) \subseteq \text{Star}_x(T) - \text{Star}_x(T)$  for any choice of  $x \in \mathbb{R}^n$ , hence

$$w(\text{Star}_0(G(X) - G(X)) \cap \rho B_2^n) \leq w((\text{Star}_x(G(X)) - \text{Star}_x(G(X))) \cap \rho B_2^n).$$



In the case of a  $\gamma$ -Lipschitz generative model, we can construct  $\text{Star}_x(G(X))$  as the image of a different Lipschitz continuous function and therefore use the results from the previous section. This construction is given in the following lemma.

**Lemma 2.4.3.** *Let  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  be  $\gamma$ -Lipschitz. Choose some  $x_0 \in X$  and define the star-shaped version of  $G$  around  $G(x_0)$ ,  $G^* : X \times [0, 1] \rightarrow \mathbb{R}^n$  as*

$$G^*(x, \lambda) := \lambda G(x) + (1 - \lambda)G(x_0).$$

*Then the image  $G^*(X, [0, 1])$  is star-shaped around the chosen point  $G(x_0)$  and  $G^*$  has Lipschitz constant  $\sqrt{2}\gamma(1 + \text{diam}(X))$ .*

*Proof.* The full proof can be found in appendix A, as it follows from a standard analysis argument, but a picture of the idea of the proof is given in Figure 2.2. ■

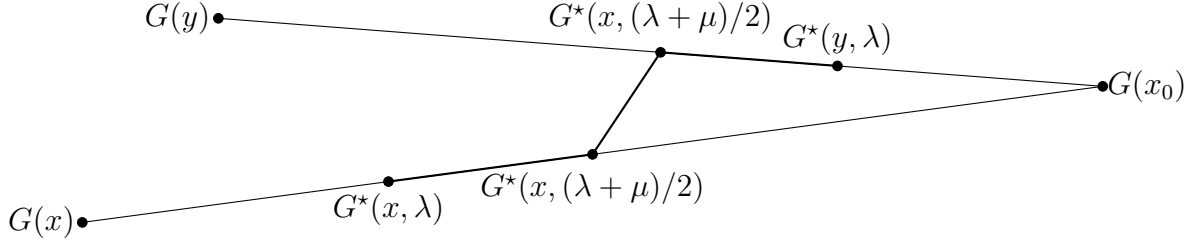


Figure 2.2: Picture of the proof of Lemma 2.4.3.

Note that the actual choice of  $x_0$  does not matter for the Lipschitz constant of  $G^*$ , therefore we let the choice be arbitrary.

In order to use the results of Lemma 2.3.2 we need to understand the covering number of  $X \times [0, 1]$ , which can be easily bounded using the following lemma.

**Lemma 2.4.4.** *If  $T \in \mathbb{R}^n$ ,  $S \in \mathbb{R}^m$  and  $\epsilon > 0$ , then*

$$\mathcal{N}(T \times S, \epsilon) \leq \mathcal{N}\left(T, \frac{\epsilon}{\sqrt{2}}\right) \mathcal{N}\left(S, \frac{\epsilon}{\sqrt{2}}\right).$$

Using Lemma 2.4.4 above we obtain

$$\mathcal{N}(X \times [0, 1], \epsilon) \leq \mathcal{N}\left(X, \frac{\epsilon}{\sqrt{2}}\right) \mathcal{N}\left([0, 1], \frac{\epsilon}{\sqrt{2}}\right) = \left\lceil \frac{\sqrt{2}}{2\epsilon} \right\rceil \mathcal{N}\left(X, \frac{\epsilon}{\sqrt{2}}\right).$$

By combining all the previously obtained lemmas we get the following bound for the localized Gaussian width of the star shaped hull of a set.

**Lemma 2.4.5.** *Let  $G : B_2^k \rightarrow \mathbb{R}^n$  be  $\gamma$ -Lipschitz, then for any  $\rho > 0$ ,*

$$w(\text{Star}_0(G(B_2^k) - G(B_2^k)) \cap \rho B_2^n) \leq C\rho \left( \sqrt{k \log \left( \frac{c\sqrt{n}\gamma}{\rho} + 1 \right)} + 1 \right),$$

where  $C, c > 1$  are universal constants.

Compared to the non-star-shaped bound

$$w((G(B_2^k) - G(B_2^k)) \cap \rho B_2^n) \leq C\rho \left( \sqrt{k \log \left( \frac{4\sqrt{n}\gamma}{\rho} + 1 \right)} + 1 \right),$$

nothing has changes besides some constants.

Most of the bounds found in this chapter for the various size and complexity quantities of the signal sets will be used in the next chapter to derive results for one-bit compressed sensing for generative models from results on one-bit compressed sensing for general signal sets.

# Chapter 3

## (Sub-)Gaussian setting

In this chapter, we will discuss various reconstruction guarantees for one-bit compressed sensing with generative models where the measurements vectors and often also the noise have Gaussian or sub-Gaussian distribution. Most of these results hold for general signal sets and will be combined with the results from the previous chapter to get results for generative models. This chapter will be concluded with a discussion of the optimality of the found sampling complexities.

### 3.1 Gaussian measurements

Plan and Vershynin [29] studied the use of convex programming to recover signals from very general measurements including one-bit quantization. Their general results allow for various models of noise, including pre-quantization noise and random bit-flips. The following theorem, which is a slight variation of a corollary from their paper.

**Theorem 3.1.1.** *[Based on Corollary 1.2 of [29]] Let  $A$  be a random  $m \times n$  matrix with i.i.d. standard Gaussian entries,  $T \subset RB_2^n$ ,  $\xi \sim N(0, \sigma^2 I)$  and fix  $t_0 \in T \cap S_2^{n-1}$ . Consider measurements of the form  $y = \text{sign}(At_0 + \xi)$  and let  $\hat{t}$  be the solution of the optimization problem*

$$\min_{t \in T} \|t\|_2^2 - \frac{2}{\lambda m} y^T A t \quad \text{with} \quad \lambda = \sqrt{\frac{2}{\pi(\sigma^2 + 1)}}. \quad (3.1)$$

*Then there exists universal constants  $c, C > 0$  such that for any  $\epsilon > 0$ , if*

$$m \geq C \frac{w^2(T)}{\lambda^2 \epsilon^2},$$

*then with probability at least  $1 - 4 \exp(-c \lambda^2 \epsilon^2 m)$ ,*

$$\|\hat{t} - t_0\|_2^2 \leq \epsilon.$$

The major differences compared to the original theorem is the added regularization that allows for arbitrary signal sets and the specific choice of one-bit quantization with Gaussian pre-quantization noise.

Also note that that in Theorem 3.1.1 the high-probability guarantee holds for a fixed true signal  $t_0$ . The other results discussed in this thesis are uniform results, meaning that the high-probability guarantee holds uniformly for any true signal  $t_0$ .

The keys to the proof of Theorem 3.1.1 are the following two lemmas regarding the correlation function

$$f_{t_0}(t) := \frac{1}{m} \sum_{i=1}^m y_i \langle a_i, t \rangle, \quad (3.2)$$

with measurements  $y_i = \text{sign}(\langle a_i, t_0 \rangle + \xi_i)$ .

**Lemma 3.1.2** (Proposition 4.2 in [29]). *For  $f$  as defined above, let  $T \subset \mathbb{R}\mathbb{R}^n$  and  $u > 0$ , then*

$$\mathbb{P} \left( \sup_{t \in T-T} |f_{t_0}(t) - \mathbb{E}f_{t_0}(t)| \geq 8w(T)/\sqrt{m} + u \right) \leq 4 \exp(-mu^2/8).$$

**Lemma 3.1.3** (Lemma 4.1 in [29]). *Let  $f_{t_0}$  as defined above. For any  $t_0 \in S^{n-1}$  and  $t \in \mathbb{R}^n$ , we have*

$$\mathbb{E}f_{t_0}(t) = \lambda \langle t_0, t \rangle,$$

with  $\lambda = \sqrt{\frac{2}{\pi(\sigma^2+1)}}$ .

*Proof of Theorem 3.1.1.* By minimality of  $\hat{t}$ , we get that

$$\|\hat{t}\|_2^2 - \frac{2}{\lambda m} y^T A \hat{t} \leq \|t_0\|_2^2 - \frac{2}{\lambda m} y^T A t_0,$$

or equivalently

$$0 \leq \|t_0\|_2^2 - \|\hat{t}\|_2^2 + \frac{2}{\lambda} f(\hat{t} - t_0).$$

Using the Lemma 3.1.2, we get for any  $u > 0$ , with probability at least  $1 - 4 \exp(-mu^2/8)$ ,

$$0 \leq \|t_0\|_2^2 - \|\hat{t}\|_2^2 + \frac{2\mathbb{E}f(\hat{t} - t_0)}{\lambda} + \frac{2}{\lambda} \left( \frac{8w(T)}{\sqrt{m}} + u \right).$$

By Lemma 3.1.3,

$$\begin{aligned} 0 &\leq \|t_0\|_2^2 - \|\hat{t}\|_2^2 + 2\langle x^*, \hat{t} - t_0 \rangle + \frac{2}{\lambda} \left( \frac{8w(T)}{\sqrt{m}} + u \right) \\ &\leq -\|\hat{t} - t_0\|_2^2 + \frac{2}{\lambda} \left( \frac{8w(T)}{\sqrt{m}} + u \right). \end{aligned}$$

Choosing  $u = c' \lambda \epsilon$  and  $m \geq C' w^2(T)/\lambda^2 \epsilon^2$  with constants  $c'$  and  $C'$  completes the proof. ■

Theorem 3.1.1 can directly be applied to the range of a generative model and combined with either Lemma 2.2.2 or 2.3.3 to get the following result.

**Corollary 3.1.4.** *Let  $A$  be a random  $m \times n$  matrix with i.i.d. standard Gaussian vectors as rows. Let  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  be  $\gamma$ -Lipschitz with  $G(X) \subseteq RB_2^n$ ,  $\xi \sim N(0, \sigma^2 I)$  and choose  $x_0 \in \mathbb{R}^k$  such that  $G(x_0) \in S_2^{n-1}$ . Consider measurements of the form  $y = \text{sign}(AG(x_0) + \xi)$  and let  $\hat{x}$  be the solution of the optimization problem*

$$\min_{x \in X} \|G(x)\|_2^2 - \frac{2}{\lambda m} y^T AG(x) \quad \text{with} \quad \lambda = \sqrt{\frac{2}{\pi(\sigma^2 + 1)}}. \quad (3.3)$$

Then there exist universal constants  $c, C > 0$  such that for any  $\epsilon > 0$ , if either

1.  $m \geq C \frac{\gamma^2 w^2(X)}{\lambda^2 \epsilon^2}$  or
2.  $m \geq C \frac{R^2}{\lambda^2 \epsilon^2} \left( \log \left( \mathcal{N} \left( X, \frac{R}{\sqrt{n}\gamma} \right) \right) + 1 \right),$

then with probability at least  $1 - 4 \exp(-c\rho^2 m)$ ,

$$\|G(\hat{x}) - G(x_0)\|_2^2 \leq \epsilon.$$

For our standard latent space  $X = B_2^k$ , the two sampling complexities in Corollary 3.1.4 above become

1.  $m \geq C \frac{\gamma^2 k}{\lambda^2 \epsilon^2}$  and
2.  $m \geq C \frac{R^2 k}{\lambda^2 \epsilon^2} \left( \log \left( \frac{2\sqrt{n}\gamma}{R} + 1 \right) + 1 \right),$

respectively.

Using the bound on the range gives an additional  $\log \left( \frac{2\sqrt{n}\gamma}{R} + 1 \right)$  term, but when  $R \ll \gamma$ , this results in a smaller sampling complexity.

For Gaussian measurements, there are also results that directly consider generative models in a paper by Liu et al.[26], as in the following result in the noiseless setting.

**Theorem 3.1.5** (Corollary 1 from [26]). *Let  $c_1, c_2 > 0$  be universal constants,  $A$  be a random  $m \times n$  matrix with i.i.d. standard Gaussian vectors as rows,  $r > 0$  and  $\epsilon \in (0, 1)$ . For a  $\gamma$ -Lipschitz generative model  $G : rB_2^k \rightarrow \mathbb{R}^n$  with  $G(rB_2^k) \subseteq S_2^{n-1}$ , if*

$$m \geq c_1 \frac{k}{\epsilon} \log \left( \frac{\gamma r}{\epsilon^2} \right),$$

then with probability at least  $1 - \exp(-c_2 \epsilon m)$  the following holds: For any  $G(x_0) \in G(rB_2^k)$  with noiseless measurements  $y = \text{sign}(AG(x_0))$  and  $G(\hat{x}) \in G(rB_2^k)$  with the same measurements, i.e.,  $\text{sign}(AG(x_0)) = \text{sign}(AG(\hat{x}))$ , we have

$$\|G(x_0) - G(\hat{x})\|_2 \leq \epsilon.$$

This theorem states that with enough random hyperplane, with high probability each cell in the tessellation of  $G(rB_2^k)$  by these hyperplanes has diameter at most  $\epsilon$ . Note that, because no pre-quantization noise is considered, we require that the signal set lies in the unit sphere, hence circumventing the loss of magnitude issue.

The pairing of  $\gamma$  and  $r$  in the theorem above is no coincidence and should always be found in the pair  $\gamma r$ , simply because multiplying by a factor  $r$  is an  $r$ -Lipschitz function, and the Lipschitz constant of the composition of two functions is the product of their Lipschitz constants.

Note that in the result, the sampling complexity depends on  $1/\epsilon$  slightly worse than linear, this becomes slightly worse than quadratic when noise is introduced in the measurements as in the following theorem.

**Theorem 3.1.6.** [Corollary 3 from [26]] *With  $A$  and  $G$  as in Theorem 3.1.5 and  $\epsilon \in (0, 1)$ . If*

$$m \geq c_1 \frac{k}{\epsilon^2} \log \left( \frac{\gamma r}{\epsilon} \right),$$

*then with probability at least  $1 - \exp(-c_2 \epsilon m)$  the following holds. For any  $G(x_0) \in G(rB_2^k)$  with noisy measurements  $\tilde{y}$  satisfying  $d_H(\text{sign}(AG(x^*)), \tilde{y}) \leq \beta_1$  and its approximate reconstruction  $G(\hat{x}) \in G(rB_2^k)$  that satisfies  $d_H(\text{sign}(AG(\hat{x})), \tilde{y}) \leq \beta_2$ , we have that*

$$d_S(G(x_0), G(\hat{x})) \leq \epsilon + \beta_1 + \beta_2,$$

*where  $d_H(a, b) := \frac{1}{m} \sum_{i \in [m]} \mathbf{1}_{\{a_i \neq b_i\}}$  is the normalized Hamming distance and  $d_S(a, b) := \frac{1}{\pi} \arccos \langle a, b \rangle$  is the geodesic distance.*

Theorem 3.1.6 allows for bit corruptions of the measurements and approximate recovery at the cost of an additional factor  $\frac{1}{\epsilon}$ . Also note the previous two theorems require that the range of the generative model must lie on the unit sphere, hence there is no quadratic dependency on the Lipschitz constant or radius of the signal set. This constraint can be somewhat relaxed using a normalization argument.

## 3.2 Sub-Gaussian

Dirksen and Mendelson derived results that generalize the measurement and noise distributions to sub-Gaussian and allow for the reconstruction of the norm using dithering as described in the introduction [12].

For the recovery problem, let  $t_0 \in T \subset \mathbb{R}^n$  be a signal set and

$$y_i = \text{sign}(\langle a_i, t_0 \rangle + \xi_i + \tau_i), \quad \text{for } i = 1, \dots, m,$$

where  $a_i$  are i.i.d. centered, isotropic and sub-Gaussian random vectors and  $\xi_i$  are i.i.d. centered sub-Gaussian random variables with variance  $\sigma^2$ , the dithering  $\tau_i$  are i.i.d uniformly distributed on  $[-\lambda, \lambda]$ . Assume that the sub-Gaussian norm of the

noise  $\xi_i$  and measurement vectors  $a_i$  is bounded by  $L$ . Consider corrupted measurements  $y_{\text{corr}}$  that satisfy

$$d_H(y_{\text{corr}}, y) \leq \beta,$$

where  $d_H$  is the normalized Hamming distance,  $\beta \in [0, 1)$  is a percentage of bit-flips and define the localized star shaped hull as

$$T_\rho^{\text{Star}} = (\text{Star}_0(T - T)) \cap \rho B_2^n.$$

Consider the optimization problem

$$\min_{t \in T} \|t\|_2^2 - \frac{2\lambda}{m} y_{\text{corr}}^T A t, \quad (3.4)$$

which is similar to optimization problem (3.3). In this scenario, the following recovery guarantee holds.

**Theorem 3.2.1.** *[Slight generalization of theorem 1.7 in [12]] There exist constants  $c_0, \dots, c_4$  that depends only on  $L$  for which the following holds. Let  $T \subseteq RB_2^n$ , fix  $\epsilon > 0$ , set*

$$\lambda \geq c_0(\sigma + R)\sqrt{\log(c_0/\epsilon)}$$

and let  $r = c_1\epsilon/\log(e\lambda/\epsilon)$ . If  $m$  and  $\beta$  satisfy

$$m \geq c_2\lambda^2 \left( \left( \frac{w(T_\epsilon^{\text{Star}})}{\epsilon^2} \right)^2 + \frac{\log \mathcal{N}(T, r)}{\epsilon^2} \right) \quad \text{and} \quad \beta\sqrt{\log(e/\beta)} = c_3\frac{\epsilon}{\lambda},$$

then with probability at least  $1 - 8\exp(-c_4m\rho^2/\lambda^2)$ , for any  $t_0 \in T$ , the solution  $\hat{t}$  of optimization problem (3.4) satisfies  $\|\hat{t} - t_0\|_2 \leq \epsilon$ .

The original formulation of the theorem requires  $T$  to be convex, but, without major modifications of the proof, this can be extended to arbitrary signal sets using the star-shaped hull, as was also mentioned in the survey of quantized compressed sensing by Dirksen [8].

A direct application of Theorem 3.2.1 to the range of a generative model together with the bounds from the previous chapter leads to the following result.

**Corollary 3.2.2.** *There exist constants  $c_0, \dots, c_4$  that depends only on  $L$  for which the following holds. Let  $G : B_2^k \rightarrow \mathbb{R}^n$  be a  $\gamma$ -Lipschitz generative model, consider the signal set  $G(B_2^k) \subseteq RB_2^n$ , fix  $\epsilon > 0$  small enough and set*

$$\lambda \geq c_0(\sigma + R)\sqrt{\log(c_0/\epsilon)}.$$

If  $m$  and  $\beta$  satisfy

$$m \geq c_1\frac{k\lambda^2}{\epsilon^2} \left( \log \left( \frac{c_2n\gamma}{\epsilon} \right) + \log \log \left( \frac{e\lambda}{\epsilon} \right) \right),$$

and

$$\beta\sqrt{\log(e/\beta)} = c_3\frac{\epsilon}{\lambda},$$

then, with probability at least  $1 - 8\exp(-c_4m\epsilon^2/\lambda^2)$ , for any  $x_0 \in B_2^k$ , the solution  $G(x^*)$  of optimization problem (3.4) satisfies  $\|G(x^*) - G(x_0)\|_2 \leq \epsilon$ .

Compared to the results derived in Corollary 3.1.4, the dependencies on  $\gamma$ ,  $\frac{1}{\epsilon}$ ,  $k$  and  $R$  are similar up to logarithmic factors. However, the generalization from Gaussian to sub-Gaussian distributions gives much more freedom on the measurement vectors.

In Chapter 5 we will take an in-depth look at the proof and results from Qiu et al. [30] which further generalizes the measurement vectors to sub-exponential distributions and allows for almost arbitrary noise distribution.

### 3.3 Optimality

Although the sampling complexities in Corollary 3.1.4 and 3.2.2 are very similar, the first measures the error in terms of the squared norm. The following result by Dirksen and Mendelson shows that the latter sampling complexity for sparse recovery is near-optimal.

**Theorem 3.3.1** (Variant of theorem 1.3 from [13]). *Let  $\nu_i$  be i.i.d. centred Gaussian random variables with variance  $\sigma^2$ , set  $A$  to be a (random) measurement matrix that satisfies, with probability at least 0.95,*

$$\|At\|_2 \leq \kappa\sqrt{m}\|t\|_2, \quad \text{for all } t \in \Sigma_{s,n}.$$

*Let  $\Psi$  be any recovery procedure such that, for every fixed  $t_0 \in \Sigma_{s,n} \cap B_2^n$ , when receiving as data the measurement matrix  $A$  and the noisy linear measurements  $((At_0)_i + \nu_i)_{i=1}^m$ ,  $\Psi$  returns  $t^*$  that satisfies  $\|t^* - t_0\|_2 \leq \epsilon$  with probability 0.9. Then, for  $\epsilon$  small enough, it holds that*

$$m \geq c\kappa^{-2}\sigma^2\frac{w^2(\Sigma_s^n \cap B_2^n)}{\epsilon^2},$$

for some constant  $c > 0$ .

The required property of  $A$  is satisfied by matrices that satisfy a restricted isometry property, like random (sub-)Gaussian matrices, see for example the book by Foucart and Rauhut [14, Definition 6.1 and Theorem 9.2].

This result shows the near-optimality of the sampling complexity of Theorem 3.2.1, including the optimal dependency on the Gaussian width, the inverse error and the variance of the noise. The dependency on the sub-Gaussian norm of the measurement vectors found in Theorem 3.2.1 is hidden in the universal constants and can therefore not be analyzed for optimality.

We can slightly rephrase the optimality result for the scenario of generative models. Assuming that the recovery procedure also works for scaled problems, we get that for



every  $\gamma > 0$  there exists a  $\gamma$ -Lipschitz function  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  such that, under the same conditions of the theorem above, the sampling complexity is at least

$$m \geq c\kappa^{-2}\sigma^2 \frac{w^2(G(X))}{\epsilon^2}.$$

The trick is to let the generative model  $G$  be a simple multiplication with  $\gamma$  and  $X = \Sigma_s^n$ . The rescaling property of the recovery procedure is needed to extend the properties outside the unit ball.

At the end of Chapter 5 we will look at another optimality result that is specifically derived for generative models.

## Chapter 4

### Linear and affine covering numbers of signal sets

The important quantities we needed in the previous chapters like the covering number and Gaussian width were enough for those results, but for the analysis of the statistical result by Qui et al. [30] in the next chapter we need another quantity.

**Definition 4.0.1.** Let  $T \subseteq \mathbb{R}^n$  and  $k \in [n]$ . Then the **linear covering number**  $\mathcal{C}_{\text{lin}}(T, k)$  (or  $\mathcal{C}(T, k)$ ) is defined as the minimal number of  $k$ -dimensional linear subspaces of  $\mathbb{R}^n$  needed to cover  $T$ .

The **affine covering number**  $\mathcal{C}_{\text{aff}}(T, k)$  is defined as the minimal number of  $k$ -dimensional affine subspaces of  $\mathbb{R}^n$  needed to cover  $T$ .

These two quantities are equivalent in the following sense.

**Lemma 4.0.2.** *If  $T \subseteq \mathbb{R}^n$  and  $k \in [n]$ , then*

$$\mathcal{C}_{\text{aff}}(T, k) \leq \mathcal{C}_{\text{lin}}(T, k) \leq \mathcal{C}_{\text{aff}}(T, k - 1).$$

*Proof.* For the lower bound, note that any linear subspace is also an affine subspace. For the upper bound, let  $\cup_{i=1}^m P_i + p_i$  be an affine covering of dimension  $k - 1$ , then  $\cup_{i=1}^m \text{span}(P_i, p_i)$  is a linear covering of dimension at most  $k$ . ■

We define both the linear and affine version, because the linear covering number simplifies the proof in the following chapter, while the affine covering number is easier to bound for some signal sets. So let us first look at the two most important examples of signal sets with finite linear/affine covering number.

#### Sparse vectors

The simplest of non-trivial examples for which the linear covering number is finite is set of  $s$ -sparse vectors. For this set  $\Sigma_s^n$  we have

$$\left(\frac{n}{s}\right)^s \leq \mathcal{C}_{\text{lin}}(\Sigma_s^n, s) = \binom{n}{s} \leq \left(\frac{en}{s}\right)^s. \quad (4.1)$$

Through rotation, a same bound holds for sparse vectors in any basis.

#### Neural networks

Qui et al. [30] considered neural networks of the form

$$G(x) := \sigma(W_l \sigma(\dots W_2 \sigma(W_1 x))),$$

where  $W_i \in \mathbb{R}^{k_i \times k_{i-1}}$  with  $k_i \leq n$  for all  $i \in [l]$ , and  $\sigma$  is the ReLU activation function. They consider as signal set the bounded range  $G(\mathbb{R}^k) \cap RB_2^n$  for which they show <sup>1</sup> that

$$\mathcal{C}_{\text{lin}}(G(\mathbb{R}^k) \cap RB_2^n, k) \leq \left(\frac{en}{k}\right)^{kl}.$$

This result can be generalized to neural networks of the form

$$G(x) := \sigma(b_l + W_l \sigma(\dots b_2 + W_2 \sigma(b_1 + W_1 x))), \quad (4.2)$$

i.e., a neural network with bias. Furthermore, instead of ReLU activation functions, we can also consider activation functions that componentwise consist of  $p$  linear pieces, like the (leaky) ReLU consists of 2 linear pieces. For such a network we have

$$\mathcal{C}_{\text{aff}}(G(\mathbb{R}^k) \cap RB_2^n, k) \leq \left(\frac{epn}{k}\right)^{kl}.$$

This generalization gives much more flexibility in the type of networks we can consider and similar results can be derived for example for Convolutional Neural Networks, for which convolution and max/average pooling operations are all piecewise linear.

It is important to note that general non-linear behaviour like the use of sigmoid and logistic activation functions cannot be efficiently handled by the linear and affine covering number, which is one of the big limitations of this approach.

## 4.1 Back to the (sub-)Gaussian setting

The linear and affine covering number can be used to bound the (localized) Gaussian width and covering number of a signal set. Therefore, we can write the sampling complexities of the previous chapter in terms of these new covering numbers. In this section, we will derive these new bounds.

The linear and affine covering numbers tell us nothing about whether the set itself is bounded, therefore we will always look at the set  $T \cap RB_2^n$  to circumvent this problem. Also, most of these results work with both the linear and affine covering number, so when the specific choice does not matter, we will use  $\mathcal{C}_{\text{lin/aff}}$  to denote either  $\mathcal{C}_{\text{lin}}$  or  $\mathcal{C}_{\text{aff}}$ .

### 4.1.1 Covering Number and (Localized) Gaussian Width

By definition of the  $k$ -dimensional linear covering number, there exists  $k$ -dimensional linear subspaces  $P_1, \dots, P_{\mathcal{C}(T,k)}$  such that  $T \subseteq \cup_{i \in [\mathcal{C}(T,k)]} P_i$ . Such a decomposition will be important in the sizable proof in the next chapter, but also for the various derivations in the remainder of this chapter, like in the following lemma.

---

<sup>1</sup>Technically, they did not use the idea of the linear covering number, but the number of linear piece the domain  $\mathbb{R}^k$  is split into by the network, which is the same. Furthermore, they also used the weaker bound  $\sum_{i=0}^k \binom{k}{d} \leq d^k + 1$  instead of bounding it by the slightly stronger  $(ed/k)^k$ . See the proof of Lemma A.2 in [30] for more details.

**Lemma 4.1.1.** *If  $T \subseteq \mathbb{R}^n$ , then for any  $R \geq 0$  and  $\epsilon > 0$ ,*

$$\mathcal{N}(T \cap RB_2^n, \epsilon) \leq \mathcal{C}_{\text{lin/aff}}(T, k) \left( \frac{4R}{\epsilon} + 1 \right)^k.$$

The proof of Lemma 4.1.1 follows directly from the bound on the covering number of the union of linear subspaces. A similar argument can be done for the Gaussian width, but it requires strong Gaussian concentration inequalities. The proof of the following lemma can be found in appendix A.

**Lemma 4.1.2.** *If  $T \subseteq \mathbb{R}^n$ , then for any  $R \geq 0$ ,*

$$w(T \cap RB_2^n) \leq CR \left( \sqrt{\log \mathcal{C}_{\text{lin/aff}}(T, k)} + \sqrt{k} \right),$$

for some constant  $C > 0$ .

To bound the localized Gaussian width, it is enough to bound the linear covering number of the star-shaped hull  $\text{Star}_0(T - T)$ . Luckily, we can combine the inequalities

$$\mathcal{C}_{\text{lin/aff}}(T \pm S, k + m) \leq \mathcal{C}_{\text{lin/aff}}(T, k) \mathcal{C}_{\text{lin/aff}}(S, m),$$

and

$$\mathcal{C}_{\text{lin}}(\text{Star}_0(T), k) \leq \mathcal{C}_{\text{lin}}(T, k),$$

to conclude that the linear covering number of  $\text{Star}_0(T - T)$  is bounded by  $\mathcal{C}_{\text{lin}}(T, k)^2$ , hence we get the following corollary of Lemma 4.1.2.

**Corollary 4.1.3.** *If  $T \subseteq \mathbb{R}^n$ , then for any  $\rho \geq 0$ ,*

$$w(\text{Star}_0(T - T) \cap \rho B_2^n) \leq C\rho \left( \sqrt{\log \mathcal{C}_{\text{lin}}(T, k)} + \sqrt{k} \right),$$

for some constant  $C > 0$ .

In many practical examples, like sparse vectors and neural networks, the linear/affine covering number term will dominate the term  $\sqrt{k}$ .

## 4.1.2 Sampling Complexities

We can now combine these bounds with the sampling complexities found in the previous chapter. Recall that the sampling complexity for Gaussian measurements as in Theorem 3.1.1 is given by

$$m \geq C \frac{w^2(T)}{\epsilon^2},$$

therefore, with the example of the sparse vectors  $T = \Sigma_s^n \cap RB_2^n$ , we get a new sampling complexity of

$$m \geq CR^2 \frac{s \log(en/s)}{\epsilon^2},$$

which is the same bound as given in the work by Plan and Vershynin [29].

More interesting is using a generative model, where we assume that our generative model is a neural network as described in Equation (4.2) with ReLU activation functions. Under this assumption, we have

$$\mathcal{C}_{\text{lin}}(G(\mathbb{R}^k) \cap RB_2^n, k) \leq \left(\frac{en}{k}\right)^{kl}.$$

The sampling complexity for sub-Gaussian measurements as in Theorem 3.2.1 then becomes

$$m \geq c_2 \frac{\lambda^2}{\epsilon^2} \left( kl \log \left( \frac{en}{k} \right) + k \log \left( \frac{5R}{\epsilon} \right) \right),$$

for small enough  $\epsilon$ . Thus we see that up to some additional logarithmic terms, the linear covering number gives the dominating term.

## 4.2 VC-Dimension

Suppose two signals  $t_1, t_2 \in T$  are close together in the sense that  $\|t_1 - t_2\|_2 \leq \delta$  for some  $\delta > 0$ . It will be useful to understand how well they are separated by the measurement vectors  $a_i$ , as signals that are close together have measurements that are also close together. We can measure this using

$$\sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t_1 - t_2 \rangle| \geq \eta\}},$$

for some  $\eta > 0$ . This process counts by how many of the measurement vectors they are badly separated. For uniform results, we would like this quantity to be low with high probability for all signals that are close together, i.e.,

$$\sup_{t \in (T-T) \cap \delta B_2^n} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t \rangle| \geq \eta\}}.$$

Understanding this quantity is an important part of the proof in the next chapter.

Let  $\mathcal{P}$  be a set of  $k$ -dimensional linear spaces that cover  $T$  such that  $|\mathcal{P}| = \mathcal{C}(T, k)$ , then we can bound this supremum as

$$\sup_{t \in (T-T) \cap \delta B_2^n} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t \rangle| \geq \eta\}} \leq \sup_{P_1, P_2 \in \mathcal{P}, t \in (P_1 - P_2) \cap B_2^n} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t \rangle| \geq \eta/\delta\}}.$$

To further bound this term with high probability, we will study the combinatorics of this process through the VC-dimension of the following sets. Let the elements of  $\mathcal{P}$  be  $P_1, \dots, P_{\mathcal{C}(T, k)}$  and  $c \geq 0$  be a fixed constant. Define the following classes of functions:

$$\begin{aligned} \mathcal{H}_{i,j} &:= \{\mathbf{1}_{\{|\langle \cdot, t \rangle| \geq c\}} : t \in (P_i - P_j) \cap B_2^n\}, \quad \text{for } i, j = 1, \dots, \mathcal{C}(T, k), \quad \text{and} \\ \mathcal{H} &:= \cup_{i,j=1}^{\mathcal{C}(T, k)} \mathcal{H}_{i,j}. \end{aligned}$$

Our interest will be in the VC-dimension of the class  $\mathcal{H}$ .

**Lemma 4.2.1.** *For the class  $\mathcal{H}$  as defined above we have*

$$VC(\mathcal{H}) \leq c_0 \log \mathcal{C}(T, k),$$

for some constant  $c_0 > 0$ .

Furthermore,

$$\mathcal{R}ad_m(\mathcal{H}) \leq c_1 \sqrt{\frac{\log \mathcal{C}(T, k)}{m}},$$

for some constant  $c_1 > 0$  and where  $\mathcal{R}ad_m$  is the empirical Rademacher complexity.

*Proof.* Let us first consider the smaller class of half-space indicators

$$\hat{\mathcal{H}}'_{i,j} := \{\mathbf{1}_{\{\langle \cdot, t \rangle \geq c\}} : t \in P_i - P_j\}.$$

Because the underlying space of functions  $a \mapsto \langle a, t \rangle - c$  is an affine vector space of dimension at most  $2k$  it holds [34, Theorem 1.9] that  $VC(\hat{\mathcal{H}}'_{i,j}) \leq 2k$ . By the Sauer-Shelah lemma [32, Theorem 8.3.16] we can therefore bound the growth function like

$$\Pi(\hat{\mathcal{H}}'_{i,j}, p) \leq \left(\frac{ep}{2k}\right)^{2k}.$$

Next we consider the two-sided class

$$\hat{\mathcal{H}}_{i,j} := \{\mathbf{1}_{\{|\langle \cdot, t \rangle| \geq c\}} : t \in P_i - P_j\}.$$

Through preservation of the growth function under complement and by taking pairwise intersection [28], we can bound its growth function by

$$\Pi(\hat{\mathcal{H}}_{i,j}, p) \leq \left(\frac{ep}{2k}\right)^{4k}.$$

Now note that  $\mathcal{H}_{i,j} \subset \hat{\mathcal{H}}_{i,j}$ , hence  $\Pi(\mathcal{H}_{i,j}, p) \leq \Pi(\hat{\mathcal{H}}_{i,j}, p)$  and through a union argument we get that

$$\Pi(\mathcal{H}, p) \leq \sum_{i,j=1}^{\mathcal{C}(T,k)} \Pi(\mathcal{H}_{i,j}, p) \leq \mathcal{C}(T, k)^2 \left(\frac{ep}{2k}\right)^{4k}. \quad (4.3)$$

We can recover a bound for the VC-dimension by finding a small  $p$  such that  $\Pi(\mathcal{H}, p) < 2^p$ , in which case  $VC(\mathcal{H}) \leq p$ . For this it is enough to find  $p$  such that

$$2 \log \mathcal{C}(T, k) + 4k \log \left(\frac{ep}{2k}\right) < p.$$

This holds for  $p = c_0 \log \mathcal{C}(T, k)$  for large enough  $c_0$ , see Lemma A.0.3, allowing us to conclude that  $VC(\mathcal{H}) \leq c_0 \log \mathcal{C}(T, k)$ .

The bound on the Rademacher complexity of  $\mathcal{H}$  follows directly from the bound on the VC-dimension together with the inequality [34, Corollary 1.21]

$$\mathcal{R}ad_m(\mathcal{H}) \leq c_1 \sqrt{\frac{VC(\mathcal{H})}{m}},$$

for some constant  $c_1 > 0$ . ■

Now define the empirical process

$$\hat{R}_m(t) := \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t \rangle| \geq \eta/\delta\}},$$

and its expectation

$$R(t) := \mathbb{E}[\hat{R}_m(t)] = \mathbb{P}(|\langle a_1, t \rangle| \geq \eta/\delta).$$

The process  $\hat{R}_m(t)$  can be uniformly bounded using the following corollary.

**Corollary 4.2.2.** *Let  $u > 0$ , then with probability at least  $1 - 2e^{-u}$ ,*

$$\sup_{P_1, P_2 \in \mathcal{P}, t \in (P_1 - P_2) \cap B_2^n} |R(t) - \hat{R}_m(t)| \leq c_2 \sqrt{\frac{\log \mathcal{C}(T, k) + u}{m}},$$

for some constant  $c_2 \geq 0$ . Furthermore, we get

$$\sup_{P_1, P_2 \in \mathcal{P}, t \in (P_1 - P_2) \cap B_2^n} \hat{R}_m(t) \leq \sup_{z \in B_2^n} R(z) + c_2 \sqrt{\frac{\log \mathcal{C}(T, k) + u}{m}}.$$

*Proof.* The first bound follows directly from the fact [34, Theorem 1.14] that for  $u > 0$ , with probability at least  $1 - 2e^{-u}$  it holds that

$$\sup_{P_1, P_2 \in \mathcal{P}, t \in (P_1 - P_2) \cap B_2^n} |R(t) - \hat{R}_m(t)| \leq 2\mathcal{R}ad_m(\mathcal{H}) + \sqrt{\frac{u}{2m}}.$$

For the second inequality, notice that

$$\sup_{P_1, P_2 \in \mathcal{P}, t \in (P_1 - P_2) \cap B_2^n} \hat{R}_m(t) - \sup_{z \in B_2^n} R(z) \leq \sup_{P_i, P_j \in \mathcal{P}: t \in (P_i - P_j) \cap B_2^n} |R(t) - \hat{R}_m(t)|$$

■

Corollary 4.2.2 above will be a key component in bounding the Hamming distance in the proof in the next chapter.

# Chapter 5

## Sub-exponential setting

The results in Chapter 3 were all based on strong (sub-)Gaussian concentration inequalities and generic chaining, therefore these results cannot be easily extended to measurement vectors with distributions with tails heavier than sub-Gaussian. In this chapter, we will focus on the statistical result by Qiu et al. [30], which is a step in going beyond sub-Gaussian measurements in one-bit compressed sensing with dithering.

Their result is specifically proven when the signal set is the range of an unbiased neural network with ReLU activation functions, but they state that their proof can be extended to other neural networks with piecewise linear behaviour. The biggest part of this chapter will therefore be dedicated to the proof of this theorem when the signal set has finite linear covering number, which is the most natural extension of their result. This does not just extend the result to other neural network with piecewise linear behaviour, but also to the standard sparsity assumption of  $\Sigma_s^n$ . Besides generalizing the signal sets allowed by the theorem, some corrections and improvements have been made to the arguments in the original proof. In the next chapter we will look into further generalizing the distributions of the noise and measurement vectors beyond sub-exponential distributions.

### 5.1 The Main Result

Let  $T \subseteq \mathbb{R}^n$  be a signal set and let the objective function  $L : T \rightarrow \mathbb{R}$  be defined by

$$L(t) := \|t\|_2^2 - \frac{2\lambda}{m} \sum_{i=1}^m y_i \langle a_i, t \rangle,$$

where  $y_i = \text{sign}(\langle a_i, t_0 \rangle + \xi_i + \tau_i)$  with  $t_0 \in T$ . Denote the minimizer of  $L$  over  $T$  as  $\hat{t}$ , then the goal is to show that for suitably chosen values of  $\lambda$  and  $m$ , we have for  $u > 0$ , that with probability at least  $1 - c_0 e^{-c_1 u}$ ,  $\hat{t}$  satisfies

$$\|\hat{t} - t_0\| \leq \epsilon.$$

This result is given in the following lemma. We will always assume that the random variables we work with are continuously distributed.



**Theorem 5.1.1** (Generalization of Theorem 3.2 in [30]). *Suppose  $a_i$  are independently sampled from a mean zero, isotropic, and sub-exponential distribution. Also assume that the noise  $\xi_i$  is independently sampled from a sub-exponential distribution. For any  $\epsilon \in (0, 1)$  and  $R \geq 1$ , set  $C_{a,R,\xi} := c_1(\|a_1\|_{\psi_1}R + \|\xi_1\|_{\psi_1})$  and assume that  $\lambda \geq c_2 C_{a,R,\xi} \log(c_3 C_{a,R,\xi}/\epsilon)$  and*

$$m \geq c_4 \frac{\lambda^2}{\epsilon^2} \log^2(\lambda m) (u + \log \mathcal{C}(T, k) + k \log(2R) + k \log(m)).$$

Then, with probability at least  $1 - c_5 e^{-c_6 u}$ ,  $\hat{t}$  satisfies

$$\|\hat{t} - t_0\|_2 \leq \epsilon,$$

for any  $t_0 \in T \cap RB_2^n$ , where  $c_1, \dots, c_6 \geq 0$  are absolute constants, depending only on  $\|a_1\|_{\psi_1}$ .

## 5.2 Bias-Variance decomposition

In order to prove Theorem 5.1.1, we will prove that if a signal is sufficiently far from the true signal, say  $\|t - t_0\|_2 > \epsilon$ , then  $L(t) - L(t_0) > 0$  and therefore  $t$  cannot be the minimizer. Therefore, the minimizer  $\hat{t}$  which satisfies  $L(\hat{t}) - L(t_0) \leq 0$ , also satisfies  $\|\hat{t} - t_0\|_2 \leq \epsilon$ .

To achieve this, we can decompose the excess objective  $L(t) - L(t_0)$  in two components and bound each component separately. Therefore, we define a deterministic bias part

$$B(t, t_0) := \|t\|_2^2 - \|t_0\|_2^2 - 2\lambda \mathbb{E}[y_1 \langle a_1, t - t_0 \rangle],$$

and a random variance part

$$V(t, t_0) := \frac{2\lambda}{m} \sum_{i=1}^m (y_i \langle a_i, t - t_0 \rangle - \mathbb{E}[y_i \langle a_i, t - t_0 \rangle]),$$

such that

$$L(t) - L(t_0) = B(t, t_0) - V(t, t_0).$$

In the next couple of sections, we will show that under conditions of Theorem 5.1.1, if  $\|t - t_0\|_2 > \epsilon$ , then  $B(t, t_0) \geq \frac{1}{2}\|t - t_0\|_2^2$  and  $V(t, t_0) \leq \frac{1}{4}\|t - t_0\|_2^2$ . From this we can conclude that if  $\|t - t_0\|_2 > \epsilon$ , then

$$L(t) - L(t_0) = B(t, t_0) - V(t, t_0) \geq \frac{1}{4}\|t - t_0\|_2^2 > 0,$$

which would conclude the proof of Theorem 5.1.1.

### 5.3 Bound for Bias

In order to lower bound the bias term, we will show that there exists a function  $K(\lambda)$  such that

$$\left| \mathbb{E}[y_1 \langle a_1, t - t_0 \rangle] - \frac{1}{\lambda} \langle t_0, t - t_0 \rangle \right| \leq K(\lambda) \|t - t_0\|_2.$$

If it is then possible for any  $\epsilon \in (0, 1)$  to choose  $\lambda$  such that  $K(\lambda) \leq \frac{\epsilon}{4\lambda}$ , then we get that if  $\|t - t_0\|_2 > \epsilon$ , then

$$\begin{aligned} B(t, t_0) &= \|t\|_2^2 - \|t_0\|_2^2 - 2\lambda \mathbb{E}[y_i \langle a_i, t - t_0 \rangle] \\ &\geq \|t\|_2^2 - \|t_0\|_2^2 - 2\langle t_0, t - t_0 \rangle - \frac{1}{2}\epsilon \|t - t_0\|_2 \\ &= \|t - t_0\|_2^2 - \frac{1}{2}\epsilon \|t - t_0\|_2 \\ &> \frac{1}{2}\|t - t_0\|_2^2. \end{aligned}$$

A function  $K(\lambda)$  can be found to work with quite general measurements and noise, only depending on the tail behaviour of the unquantized, undithered measurements.

**Lemma 5.3.1.** *Let  $\lambda > 0$  and define the independent random variables  $\tau \sim \text{Unif}([- \lambda, \lambda])$ , an isotropic, mean zero random vector  $a$  and an arbitrary random variable  $\xi$ . Then,*

$$\left| \mathbb{E}[y \langle a, t - t_0 \rangle] - \frac{1}{\lambda} \langle t_0, t - t_0 \rangle \right| \leq K(\lambda) \|t - t_0\|_2,$$

with

$$y = \text{sign}(\langle a, t_0 \rangle + \xi + \tau), \quad \text{and} \\ K(\lambda) = 2\sqrt{\mathbb{P}(|V| > \lambda) + \frac{2}{\lambda^2} \int_{\lambda}^{\infty} u \mathbb{P}(|V| > u) du},$$

where  $V := \langle a, t_0 \rangle + \xi$ .

*Proof.* Let  $V := \langle a, t_0 \rangle + \xi$  and  $Z := \langle a, t - t_0 \rangle$ . Because  $a$  is isotropic and has mean zero, we have  $\mathbb{E}[ZV] = \langle t_0, t - t_0 \rangle$ , hence

$$\left| \mathbb{E}[y \langle a, t - t_0 \rangle] - \frac{1}{\lambda} \langle t_0, t - t_0 \rangle \right| = \left| \mathbb{E}[Z \text{sign}(V + \tau)] - \frac{\mathbb{E}[ZV]}{\lambda} \right|.$$

First we consider the decomposition

$$\mathbb{E}[Z \text{sign}(V + \tau) | Z, V] = \frac{ZV}{\lambda} \mathbf{1}_{\{|V| \leq \lambda\}} + Z \mathbf{1}_{\{V > \lambda\}} - Z \mathbf{1}_{\{V < -\lambda\}},$$

whose proof can be found in Appendix A as Lemma A.0.1. From this decomposition we obtain the following bound

$$\begin{aligned}
\left| \mathbb{E}[Z \text{sign}(V + \tau)] - \frac{\mathbb{E}[ZV]}{\lambda} \right| &= \left| -\mathbb{E} \left[ \frac{ZV}{\lambda} \mathbf{1}_{\{|V| > \lambda\}} \right] + \mathbb{E} [Z \mathbf{1}_{\{V > \lambda\}}] - \mathbb{E} [Z \mathbf{1}_{\{V < -\lambda\}}] \right| \\
&\leq \mathbb{E} \left[ \left| \frac{ZV}{\lambda} \mathbf{1}_{\{|V| > \lambda\}} \right| \right] + \mathbb{E} [ |Z| \mathbf{1}_{\{|V| > \lambda\}} ] \\
&\leq 2\mathbb{E} \left[ \left| \frac{ZV}{\lambda} \mathbf{1}_{\{|V| > \lambda\}} \right| \right] \\
&\leq \frac{2}{\lambda} \|V \mathbf{1}_{\{|V| > \lambda\}}\|_{L_2} \|Z\|_{L_2}.
\end{aligned}$$

Due to the assumption that  $a$  is isotropic, we have

$$\|Z\|_{L_2} = \|t - t_0\|_2,$$

and furthermore we have

$$\|V \mathbf{1}_{\{|V| > \lambda\}}\|_{L_2}^2 \leq \lambda^2 \mathbb{P}(|V| > \lambda) + 2 \int_{\lambda}^{\infty} u \mathbb{P}(|V| > u) du.$$

A proof of this final tail bound can be found in Lemma A.0.2 in Appendix A. Combining these bounds concludes the proof. ■

From this general lemma, we can derive the bounds for more specific distributions. The following corollary results in bounds similar to that of the original work by Qiu et al. [30, Lemma A.1] in the sub-exponential case, but from the above lemma we can also derive sub-Gaussian results similar to that of Dirksen et al. [12, Lemma 4.1]

**Corollary 5.3.2.** *Let  $\lambda > 0$  and define the independent random variables  $\tau_i \sim \text{Unif}([- \lambda, \lambda])$ , an isotropic sub-exponential random vector  $a_i$  with mean zero and a sub-exponential random variable  $\xi_i$ . Then,*

$$\left| \mathbb{E}[y_1 \langle a_1, t - t_0 \rangle] - \frac{1}{\lambda} \langle t_0, t - t_0 \rangle \right| \leq K(\lambda) \|t - t_0\|_2,$$

with

$$K(\lambda) = c_1 \sqrt{1 + \frac{1}{\lambda} C_{a,R,\xi} + \frac{1}{\lambda^2} C_{a,R,\xi}^2 \exp\left(-c_2 \frac{\lambda}{C_{a,R,\xi}}\right)},$$

where  $C_{a,R,\xi} = R \|a_1\|_{\psi_1} + \|\xi_1\|_{\psi_1}$ .

Furthermore, if  $\lambda \geq c_3 C_{a,R,\xi} \log(c_4 C_{a,R,\xi} / \epsilon)$ , then  $B(t, t_0) \geq \frac{1}{2} \|t - t_0\|_2^2$ .

*Proof.* Because  $a_1$  and  $\xi_1$  are sub-exponential, we have that  $\langle a_1, t_0 \rangle + \xi_1$  is also sub-exponential with  $\|\langle a_1, t_0 \rangle + \xi_1\|_{\psi_1} \leq R \|a_1\|_{\psi_1} + \|\xi_1\|_{\psi_1} = C_{a,R,\xi}$ . Therefore we have that

$$\mathbb{P}(|\langle a_1, t_0 \rangle + \xi_1| > \lambda) \leq c_1 \exp\left(-c_2 \frac{\lambda}{C_{a,R,\xi}}\right),$$

hence

$$\begin{aligned} \int_{\lambda}^{\infty} u \mathbb{P}(|\langle a_1, t_0 \rangle + \xi_1| > u) dt &\leq c_1 \int_{\lambda}^{\infty} u \exp\left(-c_2 \frac{u}{C_{a,R,\xi}}\right) dt \\ &= c_1(\lambda C_{a,R,\xi} + C_{a,R,\xi}^2) \exp\left(-c_2 \frac{\lambda}{C_{a,R,\xi}}\right). \end{aligned}$$

Hence we can conclude that

$$K(\lambda) = c_1 \sqrt{1 + \frac{1}{\lambda} C_{a,R,\xi} + \frac{1}{\lambda^2} C_{a,R,\xi}^2} \exp\left(-c_2 \frac{\lambda}{C_{a,R,\xi}}\right),$$

works.

Now to get  $K(\lambda) \leq \epsilon/4\lambda$ , we want that

$$c_1 \sqrt{\lambda^2 + \lambda C_{a,R,\xi} + C_{a,R,\xi}^2} \exp\left(-c_2 \frac{\lambda}{C_{a,R,\xi}}\right) \leq \epsilon,$$

for which it is enough to have

$$c_1(\lambda + C_{a,R,\xi}) \exp\left(-c_2 \frac{\lambda}{C_{a,R,\xi}}\right) \leq \epsilon,$$

which holds for  $\lambda \geq c_3 C_{a,R,\xi} \log(c_4 C_{a,R,\xi}/\epsilon)$ . ■

More general results, including the sub-Gaussian scenario and heavy tailed scenarios, will be discussed in the next chapter.

## 5.4 Bound for Variance

Recall that the variance term is defined as

$$V(t, t_0) := \frac{2\lambda}{m} \sum_{i=1}^m (y_i \langle a_i, t - t_0 \rangle - \mathbb{E}[y_i \langle a_i, t - t_0 \rangle]).$$

In order to show that under the conditions of Theorem 5.1.1 and when  $\|t - t_0\|_2 > \epsilon$ , then  $V(t, t_0) \leq \frac{1}{4}\|t - t_0\|_2^2$  holds uniformly with high probability, we can look at a symmetrization of the variance term as given in the following lemma. The proof follows directly from a lemma on symmetrization for probabilities [31, Lemma 2.3.7].

**Lemma 5.4.1.** *If  $\{y_i \langle a_i, t - t_0 \rangle\}_{i \in [m]}$  are independent, mean zero stochastic processes over  $t, t_0 \in T$ , then,*

$$\begin{aligned} &\mathbb{P}\left(\sup_{t, t_0 \in T} \frac{|\frac{1}{m} \sum_{i=1}^m y_i \langle a_i, t - t_0 \rangle - \mathbb{E}[y_i \langle a_i, t - t_0 \rangle]|}{\|t - t_0\|_2} \leq x\right) \\ &\geq \mathbb{P}\left(\sup_{t, t_0 \in T} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i y_i \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} \leq \frac{x}{4}\right), \end{aligned}$$

where  $\epsilon_i$  are independent Rademacher random variables. Therefore, if with probability at least  $p$ ,

$$\sup_{t, t_0 \in T} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i y_i \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} \leq \frac{\epsilon}{32\lambda}, \quad (5.1)$$

then with probability at least  $p$ ,

$$\sup_{t, t_0 \in T} \frac{|V(t, t_0)|}{2\lambda \|t - t_0\|_2} \leq \frac{\epsilon}{8\lambda}.$$

Therefore, if  $\|t - t_0\|_2 > \epsilon$ , then with probability at least  $p$  for any  $t, t_0 \in T$  we have

$$V(t, t_0) \leq \frac{1}{4} \|t - t_0\|_2^2.$$

Therefore, the current goal is to achieve inequality (5.1) with high probability, for which we will use a covering argument. Let  $\delta > 0$  to be determined later and denote by  $\mathcal{N}$  a minimal  $\delta$ -net of  $T$  and  $\mathcal{N}(t)$  the best approximation of  $t$  in  $\mathcal{N}$ . For simplicity, denote next to  $y_i$  the new measurement  $y_i^v := \text{sign}(\langle a_i, v \rangle + \xi_i + \tau_i)$  such that  $y_i = y_i^{t_0}$ , then we can split the symmetrization as follows

$$\begin{aligned} & \sup_{t, t_0 \in T} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i y_i \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} \\ & \leq \sup_{t, t_0 \in T, v \in \mathcal{N}} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i y_i^v \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} + \sup_{t, t_0 \in T} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i (y_i^{t_0} - y_i^{\mathcal{N}(t_0)}) \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2}. \end{aligned}$$

In the following two subsections we will bound these two terms separately. The first term with the fixed signs will be bounded using another covering argument, while the second term with the sign differences will be bounded by counting the number of times that the signs will change between  $t_0$  and  $\mathcal{N}(t_0)$ , i.e., we will bound the normalized Hamming distance  $d_H(y^{t_0}, y^{\mathcal{N}(t_0)})$ .

### 5.4.1 Fixed Signs

Bounding the fixed signs term can be done using the following lemma.

**Lemma 5.4.2.** *For any  $u > 0$  and finite set  $\mathcal{N}$ , with probability at least  $1 - 2 \exp(-u + 2 \log \mathcal{C}(T, k) + 2 \log(5)k + \log |\mathcal{N}|)$  we have*

$$\sup_{t, t_0 \in T, v \in \mathcal{N}} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i y_i^v \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} \leq c \|a_1\|_{\psi_1} \left( \sqrt{\frac{u}{m}} + \frac{u}{m} \right),$$

for some universal constant  $c > 0$ .

Thus if  $m \geq c \frac{\lambda^2}{\epsilon^2} (u + \log \mathcal{C}(T, k) + k + \log |\mathcal{N}|)$  for  $\epsilon \in (0, 1)$  with  $\lambda \geq 1$ ,  $\|a_1\|_{\psi_1} \geq 1$ , then with probability at least  $1 - 2 \exp(-u)$ ,

$$\sup_{t, t_0 \in T, v \in \mathcal{N}} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i y_i^v \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} \leq \frac{\epsilon}{64\lambda},$$

for some universal constant  $c > 0$  dependent only on  $\|a_1\|_{\psi_1}$ .

*Proof.* Let us first consider fixed  $v \in \mathcal{N}$ . It then follows that, because  $\epsilon_i$  are independent of the rest of the random variables, the distribution of  $\epsilon_i y^v$  is same as  $\epsilon_i$ , hence we need to bound

$$\sup_{t, t_0 \in T} \frac{\left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, t - t_0 \rangle \right|}{\|t - t_0\|_2}.$$

Due to  $T$  being covered by  $\mathcal{C}(T, k)$  linear subspaces of dimension  $k$ , we first restrict ourselves to pairs of these linear subspaces, before using a union bound to conclude the proof. So let  $\mathcal{P}_1$  and  $\mathcal{P}_2$  be such linear subspaces. For each  $\mathcal{P}_i$ , there exists a matrix  $W_i$  such that each point in  $\mathcal{P}_i$  can be written as  $W_i s_i$ , with  $s_i \in \mathbb{R}^k$ . Therefore we can use the following sequence of equivalences:

$$\begin{aligned} \sup_{t_1 \in \mathcal{P}_1, t_2 \in \mathcal{P}_2} \frac{\left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, t_1 - t_2 \rangle \right|}{\|t_1 - t_2\|_2} &= \sup_{s_1, s_2 \in \mathbb{R}^k} \frac{\left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, W_1 s_1 - W_2 s_2 \rangle \right|}{\|W_1 s_1 - W_2 s_2\|_2} \\ &= \sup_{s \in \mathbb{R}^{2k}} \frac{\left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, W s \rangle \right|}{\|W s\|_2} \\ &= \sup_{b \in \mathcal{E}^{2k} \cap S^{n-1}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right| =: E_m, \end{aligned}$$

where  $W$  is obtained by concatenating  $W_1$  and  $-W_2$ , and  $\mathcal{E}^{2k}$  is  $2k$  dimensional subspace spanned by the columns of  $W$ .

To bound  $E_m$ , let  $\mathcal{M}$  be a minimal  $\frac{1}{2}$ -net of  $\mathcal{E}^{2k} \cap S^{n-1}$  for which  $|\mathcal{M}| \leq 5^{2k}$ , because  $\mathcal{E}^{2k} \cap S^{n-1}$  and  $S^{2k-1}$  are equivalent up to rotation. With this covering we get that

$$\begin{aligned} E_m &\leq \sup_{b \in \mathcal{M}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right| + \sup_{b \in \mathcal{E}^{2k} \cap S^{n-1}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b - \mathcal{M}(b) \rangle \right| \\ &\leq \sup_{b \in \mathcal{M}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right| + \frac{1}{2} \sup_{b \in \mathcal{E}^{2k} \cap S^{n-1}} \left| \frac{1}{m} \sum_{i=1}^m \frac{\epsilon_i \langle a_i, b - \mathcal{M}(b) \rangle}{\|b - \mathcal{M}(b)\|_2} \right| \\ &\leq \sup_{b \in \mathcal{M}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right| + \frac{1}{2} E_m. \end{aligned}$$

Thus we have that

$$E_m \leq 2 \sup_{b \in \mathcal{M}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right|.$$

For any fixed  $b \in \mathcal{M}$  and  $u \geq 0$  we have by Bernstein's inequality, Theorem B.2.2, that, with probability at least  $1 - 2e^{-u}$ ,

$$2 \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right| \leq c \|a_1\|_{\psi_1} \left( \sqrt{\frac{u}{m}} + \frac{u}{m} \right).$$

Then using a union bound on the  $\frac{1}{2}$ -net  $\mathcal{M}$  gives us that with probability at least  $1 - 2e^{-u + \log(5)2k}$ ,

$$E_m \leq c \|a_1\|_{\psi_1} \left( \sqrt{\frac{u}{m}} + \frac{u}{m} \right).$$

Use a union bound over all the  $\mathcal{C}(T, k)^2$  pairs of linear subspaces  $\mathcal{P}_1$  and  $\mathcal{P}_2$ , and a union bound over all the vectors in  $\mathcal{N}$  concludes the proof of the first statement. The second statement follows from a few substitutions.  $\blacksquare$

## 5.4.2 Sign Differences

The trick to bounding

$$\sup_{t, t_0 \in T} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i (y_i^{t_0} - y_i^{\mathcal{N}(t_0)}) \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2},$$

is to show that with high probability and uniformly,  $d_H(y^{t_0}, y^{\mathcal{N}(t_0)}) \leq \alpha \ll 1$ , as this allows us to conclude that, with high probability,

$$\sup_{t, t_0 \in T} \max_{|I| \leq \alpha m} \frac{2}{m} \sum_{i \in I} \frac{|\langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} \leq \sup_{\bar{t} \in B_2^n} \max_{|I| \leq \alpha m} \frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle|,$$

which can then be efficiently bounded using centralization and the concentration properties of  $a_i$ .

Bounding the Hamming distance directly is quite difficult, as it can be difficult to work with the sign of the measurements. To circumvent this, note that if the measurements are close together such that  $|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| < \eta$  and one of them is far away from the origin such that  $|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| \geq \eta$  for  $\eta > 0$ , then the measurements must lie on the same side of the hyperplane and thus  $\text{sign}(\langle a_i, t_0 \rangle + \xi_i + \tau_i) = \text{sign}(\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i)$ . In order to use this argument with high probability, we will use the following lemma, whose proof can be found in Appendix A, and then prove the requirements.

**Lemma 5.4.3.** *Let  $\eta > 0$  and  $\epsilon \in [0, 1/2]$ . If with probability at least  $1 - p$ ,*

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| \geq \eta\}} \leq \epsilon,$$

*and with probability at least  $1 - q$ ,*

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| < \eta\}} \leq \epsilon,$$

*then with probability at least  $1 - p - q$ ,*

$$\sup_{t_0 \in T} d_H(y^{t_0}, y^{\mathcal{N}(t_0)}) \leq 2\epsilon.$$

For the first component, we have already done most of the work during the VC-dimension argument in the previous chapter, from which we can derive the following lemma.

**Lemma 5.4.4.** *Define*

$$\delta := \frac{\eta}{\|a\|_{\psi_1}} \log(c_1 \lambda / \eta) \quad \text{and}$$

$$\eta := c_2(\lambda + \|a\|_{\psi_1}) \sqrt{\frac{\log \mathcal{C}(T, k) + u'}{m}}.$$

If  $u' \geq u \geq 0$  and  $\mathcal{N}$  is a minimal  $\delta$ -net of  $T$ , then with probability at least  $1 - 2e^{-u}$ ,

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| \geq \eta\}} \leq 2 \frac{\eta}{\lambda}.$$

*Proof.* Let  $\mathcal{P}$  be a minimal set of  $k$ -dimensional linear subspaces that cover  $T$ . By the VC-dimension argument in Corollary 4.2.2, we get with probability at least  $1 - 2e^{-u}$ , that

$$\begin{aligned} \sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| \geq \eta\}} &\leq \sup_{P_1, P_2 \in \mathcal{P}, t \in (P_1 - P_2) \cap \delta B_2^n} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t \rangle| \geq \eta\}} \\ &= \sup_{P_1, P_2 \in \mathcal{P}, t \in (P_1 - P_2) \cap B_2^n} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t \rangle| \geq \eta / \delta\}} \\ &\leq \sup_{z \in B_2^n} \mathbb{P}(|\langle a_1, z \rangle| \geq \eta / \delta) + c_2 \sqrt{\frac{\log \mathcal{C}(T, k) + u}{m}}. \end{aligned}$$

Because  $a_i$  is sub-exponentially distributed,

$$\sup_{z \in B_2^n} \mathbb{P}(|\langle a_1, z \rangle| \geq \eta / \delta) \leq c_1 \exp\left(-\frac{\eta}{\delta \|a\|_{\psi_1}}\right) = \frac{\eta}{\lambda}, \quad (5.2)$$

where the second step is by the choice of  $\delta$ . The result then follows from the definition of  $\eta$ , because

$$c_2 \sqrt{\frac{\log \mathcal{C}(T, k) + u}{m}} \leq c_2 \frac{\lambda + \|a\|_{\psi_1}}{\lambda} \sqrt{\frac{\log \mathcal{C}(T, k) + u'}{m}} = \frac{\eta}{\lambda}.$$

■

The second term can be obtained through a covering argument.

**Lemma 5.4.5.** *Define*

$$\delta := \frac{\eta}{\|a\|_{\psi_1}} \log(c_1 \lambda / \eta),$$

$$\eta := c_2(\lambda + \|a\|_{\psi_1}) \sqrt{\frac{\log \mathcal{C}(T, k) + u'}{m}},$$



and assume that  $m \geq \log \mathcal{C}(T, k) + k \log(5R) + u'$  with  $u'$  satisfying

$$u' = u + \log \mathcal{C}(T, k) + k \log(5R) + Ck \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right).$$

and  $c_2 \geq 3$ . Then with probability at least  $1 - \exp(-c_0 u)$ ,

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| < \eta\}} \leq 2 \frac{\eta}{\lambda},$$

*Proof.* First consider a single  $\mathcal{N}(t_0)$  and  $a_i$ , then

$$\mathbb{P}(|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| < \eta) = \mathbb{P}(-\eta < (\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i) + \tau_i < \eta) \leq \frac{\eta}{\lambda}, \quad (5.3)$$

independent of the distribution of  $\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i$ . By using the Chernoff bound, Theorem B.4.1, we get with probability at least  $1 - \exp(-\eta m / 3\lambda)$  that

$$\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| < \eta\}} \leq 2 \frac{\eta}{\lambda}.$$

By a union bound over  $\mathcal{N}(T, \delta)$  this holds uniformly over the whole covering with probability at least  $1 - \exp(\log \mathcal{N}(T, \delta) - \eta m / 3\lambda)$ . The next goal is to bound this probability by the simple  $1 - e^{-cu}$ , for which we need to analyze both terms in the exponent. First note that

$$\log \left( \frac{1}{\delta} \right) \leq C \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right).$$

Hence for the metric entropy we have that

$$\log \mathcal{N}(T, \delta) \leq \log \mathcal{C}(T, k) + k \log \left( \frac{5R}{\delta} \right) \quad (5.4)$$

$$\leq \log \mathcal{C}(T, k) + k \log(5R) + Ck \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right). \quad (5.5)$$

Next, note that by definition of  $\eta$ ,

$$\frac{\eta m}{3\lambda} \geq \frac{c_2}{3} \sqrt{(\log \mathcal{C}(T, k) + u') m}.$$

By filling in the defining property of  $u'$ , this is further lower bounded by

$$\frac{c_2}{3} \sqrt{m} \sqrt{u + \log \mathcal{C}(T, k) + k \log(5R) + Ck \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right)}.$$

Furthermore, by constraint on  $m$  and definition of  $u'$ , we can rewrite the lower bound to become

$$\frac{c_2}{3} \left( u + \log \mathcal{C}(T, k) + k \log(5R) + Ck \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right) \right).$$

Therefore we have that

$$\log \mathcal{N}(T, \delta) - \eta m / 3\lambda \leq -\frac{c_2}{3} u$$

from which we can conclude that

$$1 - \exp(\log \mathcal{N}(T, \delta) - \eta m / 3\lambda) \geq 1 - \exp(-c_0 u),$$

for some universal constant  $c_0 > 0$ , concluding the proof. ■

We now have everything to finish up bounding the Hamming distance with high probability.

**Lemma 5.4.6.** *Define*

$$\begin{aligned} \delta &:= \frac{\eta}{\|a\|_{\psi_1}} \log(c_1 \lambda / \eta), \quad \text{and} \\ \eta &:= (\lambda + \|a\|_{\psi_1}) c_2 \sqrt{\frac{\log \mathcal{C}(T, k) + u'}{m}} \end{aligned}$$

and assume that  $m \geq c_3 (\mathcal{C}(T, k) + k \log(5R) + u')$  with  $u'$  satisfying

$$u' = u + \log \mathcal{C}(T, k) + k \log(5R) + Ck \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right),$$

$c_2 \geq 3$  and  $\lambda \geq \|a\|_{\psi_1}$ . Then with probability at least  $1 - 2e^{-u} - e^{-c_0 u}$ ,

$$\sup_{t_0 \in T} d_H(y^{t_0}, y^{\mathcal{N}(t_0)}) \leq 4 \frac{\eta}{\lambda},$$

where  $c_0, c_3 > 0$  are universal constants.

*Proof.* Although the whole lemma follows almost directly from combining Lemma 5.4.4 and 5.4.5 with Lemma 5.4.3, the one constraint that is worth mentioning is that  $\frac{\eta}{\lambda} \leq \frac{1}{4}$  must hold. However, this is equivalent to

$$m \geq 16c_2^2 \left( \frac{\lambda + \|a\|_{\psi_1}}{\lambda} \right)^2 (\log \mathcal{C}(T, k) + u').$$

Therefore, by the assumption that  $\lambda \geq \|a\|_{\psi_1}$ , this results in the additional universal constant  $c_3$  in the lower bound on  $m$ . ■

Under the conditions of this lemma we have with probability at least  $1 - 2e^{-u} - e^{-c_0u}$  that

$$\sup_{t, t_0 \in T} \frac{|\frac{1}{m} \sum_{i=1}^m \epsilon_i (y_i^{t_0} - y_i^{N(t_0)}) \langle a_i, t - t_0 \rangle|}{\|t - t_0\|_2} \leq \sup_{P_1, P_2 \in \mathcal{P}, \bar{t} \in (P_1 - P_2) \cap B_2^n} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle|.$$

To further bound this, we consider the following further decomposition.

$$\begin{aligned} & \sup_{\substack{P_1, P_2 \in \mathcal{P} \\ \bar{t} \in (P_1 - P_2) \cap S^{n-1}}} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| \\ & \leq \sup_{\substack{P_1, P_2 \in \mathcal{P} \\ \bar{t} \in (P_1 - P_2) \cap S^{n-1}}} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \\ & \quad + \sup_{\bar{t} \in S^{n-1}} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} \mathbb{E} |\langle a_i, \bar{t} \rangle|. \end{aligned}$$

To finish the proof of the variance, we now only need to bound these two terms.

**Lemma 5.4.7.** *If  $m$  satisfies*

$$m \geq c_2 \|a_1\|_{\psi_1}^2 \frac{\lambda^2}{\epsilon^2} \log^2(\lambda m) (u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m))$$

*then, with probability at least  $1 - 2e^{-u}$ ,*

$$\sup_{P_1, P_2 \in \mathcal{P}, \bar{t} \in (P_1 - P_2) \cap S^{n-1}} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \leq \frac{\epsilon}{128\lambda}.$$

*Proof.* Due to  $a_i$  being isotropic and sub-exponential, and  $\|\bar{t}\|_2 \leq 1$  it holds that  $|\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle|$  is also sub-exponential with sub-exponential norm at most  $2\|a\|_{\psi_1}$ . Hence, by Bernstein's inequality, Theorem B.2.2, we have with probability at least  $1 - 2e^{-u'}$ , that

$$\frac{1}{|I|} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \leq c \|a\|_{\psi_1} \left( \sqrt{\frac{u'}{|I|}} + \frac{u'}{|I|} \right),$$

hence also

$$\frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \leq c \frac{\|a\|_{\psi_1}}{m} \left( \sqrt{u'|I|} + u' \right)$$

for fixed choice of  $\bar{t}$  and  $I$ . Now let

$$u' = c \log(\lambda m) \sqrt{u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m)} \sqrt{m},$$

and note that

$$|I| \leq c\sqrt{u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m)}\sqrt{m}.$$

Replacing  $u$  with  $u'$  results in that with probability at least

$$1 - 2 \exp\left(-c \log(\lambda m) \sqrt{u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m)}\sqrt{m}\right)$$

it holds that

$$\frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E}|\langle a_i, \bar{t} \rangle| \leq c \|a\|_{\psi_1} \log(\lambda m) \sqrt{\frac{u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m)}{m}}.$$

Now by a union bound over all possible  $I$  together with an argument similar to Lemma 5.4.2, by noting that

$$\log\left(\sum_{i=0}^{\lfloor 4\eta m/\lambda \rfloor} \binom{m}{i}\right) \leq c \log(\lambda m) \sqrt{u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m)}\sqrt{m},$$

and by assumption on  $m$ , we get that the above inequality holds uniformly with probability at least

$$\begin{aligned} & 1 - 2 \exp\left(-c_1 \log(\lambda m) \sqrt{u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m)}\sqrt{m}\right. \\ & \quad \left. + 2k \log(5R) + 2 \log \mathcal{C}(T, k)\right) \\ & \geq 1 - 2 \exp(-c_1 u + k \log(5R) + 2 \log \mathcal{C}(T, k)). \end{aligned}$$

By substituting  $u$  we get with probability at least  $1 - 2 \exp(-u)$  that

$$\begin{aligned} & \sup_{P_1, P_2 \in \mathcal{P}, \bar{t} \in (P_1 - P_2) \cap S_2^n} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E}|\langle a_i, \bar{t} \rangle| \\ & \leq c \|a\|_{\psi_1} \log(\lambda m) \sqrt{\frac{u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m)}{m}}. \end{aligned}$$

The remainder of the proof follows directly from the assumption on  $m$ . ■

**Lemma 5.4.8.** *If  $m$  satisfies*

$$m \geq c \frac{(\lambda + \|a\|_{\psi_1})^2}{\epsilon^2} (u + \log \mathcal{C}(T, k) + k \log(5R) + Ck \log(m)) / \epsilon^2,$$

*then, with probability at least  $1 - c_1 e^{-u}$ ,*

$$\sup_{\bar{t} \in B_2^n} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} \mathbb{E}|\langle a_i, \bar{t} \rangle| \leq \frac{\epsilon}{128\lambda}.$$

*Proof.* Due to  $a_i$  being isotropic and  $\|\bar{t}\|_2 \leq 1$ , we have that  $\mathbb{E}|\langle a_i, \bar{t} \rangle| \leq 1$ , hence

$$\sup_{\bar{t} \in B_2^n} \max_{|I| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in I} \mathbb{E}|\langle a_i, \bar{t} \rangle| \leq \max_{|I| \leq 4m\eta/\lambda} \frac{2|I|}{m} \leq \frac{8\eta}{\lambda}.$$

Furthermore, by choice of  $\eta$  and  $u'$ , we get that

$$\frac{8\eta}{\lambda} \leq 8 \frac{\lambda + \|a\|_{\psi_1}}{\lambda} c_2 \sqrt{\frac{u + \log \mathcal{C}(T, k) + k \log(5R) + Ck \log(m)}{m}}.$$

The remainder of the proof follows directly from the constraint on  $m$ . ■

### 5.4.3 Completing the Proof

Now that all the terms are bounded, we only have to combine everything with a union bound and choose a sampling complexity that satisfies all constraints.

Because all these probabilities are of the form  $1 - c_0 e^{-c_1 u}$ , we get a final probability of the form  $1 - c_3 e^{-c_4 u}$ .

Because of the bias bound we require that  $\lambda \geq c_2 C_{a,R,\xi} \log(c_3 C_{a,R,\xi}/\epsilon)$  with  $C_{a,R,\xi} := c_1 (\|a_1\|_{\psi_1} R + \|\xi_1\|_{\psi_1})$ . As for  $m$ , it is enough for  $m$  to satisfy the dominating sampling complexity of Lemma 5.4.7. Hence

$$m \geq c_2 \frac{\lambda^2}{\epsilon^2} \log^2(\lambda m) (u + \log \mathcal{C}(T, k) + k \log(5R) + k \log(m))$$

is enough to satisfy all of the previous requirements on  $m$ . Therefore, concluding the proof of Theorem 5.1.1.

## 5.5 Comparison

If we combine Theorem 3.2.1 with Corollary 4.1.3 and Lemma 4.1.1, we obtain in the sub-Gaussian settings the sampling complexity

$$m \geq C \frac{\lambda^2}{\epsilon^2} \left( \log \mathcal{C}(T, k) + k \log \left( \frac{4R}{\epsilon} \right) \right)$$

with  $\lambda \geq c_0 (\sigma + R) \sqrt{\log(c_0/\epsilon)}$ .

For comparison, the sampling complexity in the sub-exponential setting of Theorem 5.1.1 is

$$m \geq c_4 \frac{\lambda^2}{\epsilon^2} \log^2(\lambda m) (u + \log \mathcal{C}(T, k) + k \log(2R) + k \log(m)),$$

with  $\lambda \geq c_2 C_{a,R,\xi} \log(c_3 C_{a,R,\xi}^2/\epsilon)$  and  $C_{a,R,\xi} := c_1 (\|a\|_{\psi_1} R + \|\xi\|_{\psi_1})$ .

To simplify this, the additional logarithmic terms in  $m$  can be enforced when

$$m \geq c_4 \frac{\lambda^2}{\epsilon^2} \log^5 \left( \frac{\lambda}{\epsilon} \right) \left( u + \log \mathcal{C}(T, k) + k \log \left( \frac{2R}{\epsilon} \right) \right).$$

Therefore, the major difference between these sub-Gaussian and sub-exponential results are some addition logarithmic dependencies on  $\lambda$  and  $1/\epsilon$ .

## 5.6 Optimality

In Chapter 3 we discussed an optimality result that relied on sparse vectors, but lower bounds specifically for generative models do exist. To mention one result in particular, Qiu et al. [30] proved, based on a similar result by Liu et al. [27], the following theorem.

**Theorem 5.6.1** (Theorem 3.4 from [30]). *Let  $k$  and  $n$  be large enough and satisfying  $k \ll n$ , then there exists a ReLU neural network  $G : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^n$ , consisting of 3 layers, for which the following holds.*

*Consider the signal set  $T = G(\mathbb{R}^k) \cap B_2^n$ , with unquantized measurements  $y_i = \langle a_i, t_0 \rangle + \xi_i$  with  $t_0 \in T$  and independent random variables  $\xi \sim N(0, 1)$  and let the measurement vectors  $a_i$  composed of i.i.d. standard normal entries. If  $m \geq c_1 k \log(n/k)$ , then*

$$\sup_{t_0 \in T} \mathbb{E} \|\hat{t} - t_0\|_2 \geq c_2 \sqrt{\frac{k \log(n/k)}{m}},$$

*for any estimator  $\theta_0$  that depends only on  $a_i$  and  $y_i$  for  $i \in [m]$ , where  $c_1, c_2 > 0$  are universal constants.*

Considering that for the generative model considered in the theorem we have the bound

$$\log \mathcal{C}_{\text{lin}}(G(\mathbb{R}^k) \cap RB_2^n, k) \leq 3k \log \left( \frac{en}{k} \right),$$

which can be arbitrarily tight for large  $n$  and  $k$ , the logarithmic dependency on the linear covering number and the quadratic dependency on  $1/\epsilon^2$  seem to be optimal, and the two theorems using dithering discussed in this thesis are near optimal up to additional logarithmic factors. As for the term  $k \log(R)$ , this one is generally dominated by  $\log \mathcal{C}(T, k)$  and is therefore of little importance.

# Chapter 6

## Heavier tailed setting

Because the proof in the previous chapter does not rely on strong sub-Gaussian properties like generic chaining, it was able to generalize the permissible distributions to sub-exponential distributions by using the still very strong Bernstein's inequality, Chernoff bound and the VC-dimension argument. In this chapter we will look into ways to extend the results to heavy-tailed distributions.

### 6.1 Bias and Noise

A possibly surprising observation in the proof of the previous chapter is that in the bound on the variance, the only time the additive noise  $\xi_i$  has been used was in inequality (5.3) and there the result was independent of the actual distribution of  $\langle a_i, t \rangle + \xi_i$ , as long as it is continuously distributed. Therefore, we have the option to consider heavy tailed noise distributions, which can work with Lemma 5.3.1.

Furthermore, we can generalize the way the noise interacts with measurement. For example, a result similar to that of Lemma 5.3.1 also holds for measurements of the form  $\xi_i \langle a_i, t \rangle$ , i.e., multiplicative noise. In this section we will look at various results for heavy tailed additive and how we can allow for multiplicative noise, still under the assumption that the measurement vectors are sub-exponential.

#### 6.1.1 Sub-Weibull additive noise

Sub-exponential and sub-Gaussian random variables are both characterized by their tail behaviour, decaying like  $2e^{-cu}$  and  $2e^{-cu^2}$  respectively for some constant  $c > 0$ . In more generality, we say that a random variable is **sub-Weibull** with tail parameter  $\theta$  if the tail decays like  $2e^{-cu^{1/\theta}}$ , for some  $c > 0$ . This more general family of random variables allows for the generalization of various well known properties of sub-exponential and sub-Gaussian random variables and has various applications in high-dimensional probability [23].

Sadly, the tail integral

$$\int_{\lambda}^{\infty} u \mathbb{P}(|\langle a_1, t \rangle + \xi_1| > u) du = 2 \int_{\lambda}^{\infty} u e^{-cu^{1/\theta}} du,$$

found in Lemma 5.3.1 has no simple solution for general  $\theta > 0$ . For the easily computable sub-exponential distribution ( $\theta = 1$ ) however, we have seen in

Corollary 5.3.2 that

$$\lambda \geq c_0 C_{a,R,\xi} \log(c_1 C_{a,R,\xi}/\epsilon),$$

with  $C_{a,R,\xi} = R\|a_1\|_{\psi_1} + \|\xi_1\|_{\psi_1}$  is enough to bound the bias.

Similar computations for sub-Gaussian distributions ( $\theta = 1/2$ ) show that

$$\lambda \geq c_0 C_{a,R,\xi} \sqrt{\log(c_1 C_{a,R,\xi}/\epsilon)},$$

with  $C_{a,R,\xi} = R\|a_1\|_{\psi_2} + \|\xi_1\|_{\psi_2}$  is enough to bound the bias.

### 6.1.2 $L_p$ Additive noise

To bound the bias term when the unquantized measurements have finite  $L_p$  norm, we can again make use of Lemma 5.3.1. Replacing the exponentially decaying tail with Markov's inequality results in the following lemma, similar to Corollary 5.3.2.

**Lemma 6.1.1.** *Let  $\lambda > 0$  and define the following independent random variables  $\tau_i \sim \text{Unif}([- \lambda, \lambda])$ ,  $a_i$  be an isotropic random variable and  $\xi_i$  be an arbitrary random variable with zero mean. Furthermore, let  $\|\langle a_1, t_0 \rangle + \xi_1\|_{L_p}^p =: M < \infty$  for  $p > 2$ . Then,*

$$\left| \mathbb{E}[y_1 \langle a_1, t - t_0 \rangle] - \frac{1}{\lambda} \langle t_0, t - t_0 \rangle \right| \leq K(\lambda) \|t - t_0\|_2,$$

with

$$K(\lambda) = \frac{4}{\lambda^{p/2}} \sqrt{M + \frac{2}{p-2}}.$$

Furthermore, if  $\lambda \geq C_{p,M} (1/\epsilon)^{\frac{1}{p/2-1}}$ , then  $B(t, t_0) \geq \frac{1}{2} \|t - t_0\|_2^2$ , where  $C_{p,M}$  is a constant that only depends on  $p$  and  $M$ .

When the measurement vectors  $a_i$  are sub-exponential, the  $L_p$ -norm can be bounded like  $\|\langle a_1, t_0 \rangle + \xi_1\|_{L_p}^p \leq (Cp\|a_1\|_{\psi_1} + \|\xi_1\|_{L_p})^p$ .

For comparison, the lower bound on  $\lambda$  for sub-exponential  $\langle a_i, t_0 \rangle + \xi_i$  is logarithmic in  $\log(1/\epsilon)$  and in the sub-Gaussian case it behaves like  $\sqrt{\log(1/\epsilon)}$ , both of which are asymptotically better than the power law behaviour of the lemma above.

### 6.1.3 Multiplicative noise

Now consider the case of multiplicative noise, i.e.,

$$V = \xi_1 \langle a_1, t_0 \rangle.$$

To generalize Lemma 5.3.1 to this scenario, we need to assume that  $\mathbb{E}[\xi_i] \neq 0$  and  $\xi_i$  is independent of  $a_i$ . Furthermore, assume for simplicity that  $\mathbb{E}[\xi_i] = 1$ , otherwise we need to modify the optimization problem, and therefore the variance to handle this



scaling. Under these additional assumptions, the same result as Lemma 5.3.1 holds, because

$$\langle t_0, t - t_0 \rangle = \mathbb{E} \xi_1 \langle a_1, t_0 \rangle \langle a_1, t - t_0 \rangle$$

Therefore, based on whether  $V$  is sub-Weibull or has finite  $L_p$  norm we get similar results like in the previous two subsections, with similar asymptotic behaviour in  $1/\epsilon$ , but different dependence on the  $\psi_\theta$ -or  $L_p$ -norms.

There is a caveat however, although multiplicative noise is permissible in inequality (5.3), we would require that  $|\xi_i \langle a_i, \mathcal{N}(t_0) \rangle + \tau_i| > \eta$  and  $|\xi_i \langle a_i, t_0 - \mathcal{N}(t_0) \rangle| \leq \eta$  for  $\text{sign}(\xi_i \langle a_i, \mathcal{N}(t_0) \rangle + \tau_i) = \text{sign}(\xi_i \langle a_i, t_0 \rangle + \tau_i)$ . Lemma 5.4.5 was distribution independent, hence can be directly modified to cover the first term. Lemma 5.4.4 requires some more modifications. For example, if  $\xi_1$  and  $a_1$  are a sub-Gaussian random variable and random vector respectively, then for any  $z \in B_2^n$ ,

$$\|\xi_1 \langle a_1, z \rangle\|_{\psi_1} \leq \|\xi_1\|_{\psi_2} \|\langle a_1, z \rangle\|_{\psi_2} \leq \|\xi_1\|_{\psi_2} \|a_1\|_{\psi_2},$$

hence

$$\sup_{z \in B_2^n} \mathbb{P}(|\xi_1 \langle a_1, z \rangle| \geq \eta/\delta) \leq c_1 \exp\left(-\frac{\eta}{\delta \|\xi_1\|_{\psi_2} \|a_1\|_{\psi_2}}\right).$$

Therefore, in this sub-Gaussian setting, the difference in the Hamming distance argument becomes the term  $\|\xi_1\|_{\psi_2} \|a_1\|_{\psi_2}$  replacing  $\|a_1\|_{\psi_1}$ .

## 6.2 Heavy-tailed measurement vectors

In this section, we will discuss how to possibly generalize the measurement vectors beyond sub-exponential distributions. First, we will see that to use the argumentation of the previous chapter, inverse polynomially decaying tails is not enough for reasonable sampling complexities. Second, we will use that the tails of some random variables partially behave sub-Gaussian. This will nearly allow us to extend the results of the previous chapter to heavy tailed distribution, but as we will see, this argumentation currently causes a small, but currently unsolved, contradiction.

### 6.2.1 Convergence speed of weak law of large numbers

Bernstein's inequality shows that the convergence speed of the weak law of large numbers is exponential. To take this assumption to the extreme, we will assume in this section that the weak law of large numbers converges like a power law, more precisely, assume that there exist  $M, \alpha, \beta > 0$  such that

$$\mathbb{P}\left(\left|\frac{1}{m} \sum_{i=1}^m \epsilon \langle a_i, t \rangle\right| \geq u\right) \leq \frac{M}{m^\alpha u^\beta},$$

for all  $t \in B_2^n$  and  $t > 0$ . The easiest example of this is the case of  $\alpha = 1$  and  $\beta = 2$ , which holds if the variance of  $\epsilon \langle a_i, t \rangle$  is uniformly bounded.

We will not discuss a full sampling complexity, but only that the natural flow of the proof of last chapter results in practically unachievable, if not impossible, sampling complexities. For this we will only look at the fixed signs and Hamming distance components of the proof. Bounding the fixed signs with the weaker convergence, just like in the previous chapter, results in the following lemma.

**Lemma 6.2.1.** *For any  $u > 0$  and finite set  $\mathcal{N} \subseteq T$ , if*

$$m \geq C \left( \frac{\lambda}{\epsilon} \right)^{\beta/\alpha} (uM\mathcal{C}(T, k)^2 5^{2k} |\mathcal{N}|)^{1/\alpha}$$

for  $\epsilon \in (0, 1)$  with  $\lambda \geq 1$ , then with probability at least  $1 - \frac{1}{u}$  it holds that

$$\sup_{t, t_0 \in T, v \in \mathcal{N}} \frac{\left| \frac{1}{m} \sum_{i=1}^m \epsilon_i y_i^v \langle a_i, t - t_0 \rangle \right|}{\|t - t_0\|_2} \leq \frac{\epsilon}{64\lambda},$$

for some universal constant  $C > 0$ .

If we assume that the Hamming distance argument does not change, which is not the case, then by Equation (5.4), the set  $\mathcal{N}$  would satisfy

$$\mathcal{N}(T, \delta) \leq \mathcal{C}(T, k)(5R)^k m^{ck}.$$

If you think that the sampling complexity Lemma 6.2.1 already seems large, we now would have that  $m \geq Cm^{ck/\alpha}$ , which is impossible to achieve if  $ck \geq \alpha$ .

This argument shows just how important the exponential behaviour in Bernstein's inequality is to obtaining reasonable sampling complexities in the proof of the previous chapter. The remainder of this chapter is concerned with trying to find a way between the good exponential decay and bad polynomial decay.

## 6.2.2 Partially sub-Gaussian random variables

An important characterizing property of sub-Weibull distributions [33] is that, for some tail parameter  $\theta > 0$ ,

$$\mathbb{P}(|X| \geq u) \leq 2 \exp(-u/C_1)^{1/\theta},$$

for some constant  $C_1 > 0$  if and only if

$$\|X\|_{L^p} \leq C_2 p^\theta \quad \text{for all } p \geq 1,$$

for some constant  $C_2 > 0$ . However, some random variables do not even have moments of all order but behave similarly to sub-Weibull distributions. For example, a Student's t-distribution with  $d$  degrees of freedom only has moments up to order  $d - 1$ , yet the moments of a Student's t-distribution behave similar to sub-Gaussian random variables [21, 10].

For such a random variable, whose moments partially behave like a sub-Weibull random variable, we can show that their tails behave similar to that of a sub-Weibull random variable on some bounded set.

**Lemma 6.2.2** (Based on lemma 11 from [11]). *Let  $0 < p_0 < p_1 \leq \infty$  and  $\theta > 0$ . If  $X$  is a random variable that satisfies*

$$\|X\|_{L^p} \leq Kp^\theta,$$

*for some  $K > 0$  and for all  $p \in [p_0, p_1]$ , then*

$$\mathbb{P}(|X| \geq eKu^\theta) \leq e^{-u},$$

*for all  $u \in [p_0, p_1]$ .*

*Proof.* Let  $u \in [p_0, p_1]$ , then by Markov's inequality

$$\mathbb{P}(|X| \geq eKu^\theta) \leq \left( \frac{\|X\|_{L^u}}{eKu^\theta} \right)^u \leq \left( \frac{1}{e} \right)^u = e^{-u}.$$

■

To replace Bernstein's inequality with a heavy tailed concentration inequality, we need to understand the behaviour of sums of heavy tailed random variables. The following lemma shows that the weak moment assumption in Lemma 6.2.1 above can result in sub-Gaussian moments for the sum of random variables, therefore resulting in a partially sub-Gaussian concentration inequality for sums of enough heavy tailed random variables.

**Lemma 6.2.3** (Lemma 2.8 from [24]). *For  $m \in \mathbb{N}$ , Let  $X_1, \dots, X_m$  be i.i.d. copies of a mean zero random variable  $X$ , that for some  $p_0 \geq 2$  satisfies*

$$\|X\|_{L^p} \leq Kp^\theta, \quad \text{for all } p \in [2, p_0],$$

*for some constant  $K > 0$  and  $\theta \geq 1/2$ . If  $m \geq p_0^{\max\{2\theta-1, 1\}}$ , then*

$$\left\| \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i \right\|_{L^p} \leq C_\theta K \sqrt{p}, \quad \text{for all } p \in [2, p_0],$$

*where  $C_\theta = Ce^{2\theta-1}$  for some constant  $C > 0$ .*

Applying Lemma 6.2.2 to the Lemma 6.2.3 gives us the following partial sub-Gaussian concentration inequality for random variables that satisfy a weak moment assumption.

**Corollary 6.2.4.** *Under the conditions of Lemma 6.2.3 we have*

$$\mathbb{P} \left( \left| \frac{1}{m} \sum_{i=1}^m X_i \right| \geq eC_\theta K \sqrt{\frac{u}{m}} \right) \leq e^{-u},$$

*for  $u \in [2, p_0]$ .*

In the following sub-sections, we will try to derive recovery guarantees for one-bit compressed sensing with heavy tailed distributions whose moments partially behave like a sub-Weibull distribution.

### 6.2.3 Variance Bound - Fixed Signs

Let us first consider Lemma 5.4.2, which is the first step in the variance bound of the previous chapter where the sub-exponentially of the measurement vectors was used when using Bernstein's inequality. Specifically, it was used to show with probability at least  $1 - 2e^{-u}$  that

$$2 \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right| \leq c \|a_1\|_{\psi_1} \left( \sqrt{\frac{u}{m}} + \frac{u}{m} \right),$$

where  $b \in S^{n-1}$ .

If we assume that there exists  $p_0 \geq 2$ ,  $\theta \geq 1/2$  and  $K > 0$  such that for all  $p \in [2, p_0]$ ,

$$\sup_{x \in S^{n-1}} \|\langle a_i, x \rangle\|_{L^p} \leq K p^\theta,$$

then we can replace Bernstein's inequality with Corollary 6.2.4 and get with probability at least  $1 - e^{-u}$  that

$$2 \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, b \rangle \right| \leq 2eC_\theta K \sqrt{\frac{u}{m}},$$

with the additional constraints that  $m \geq p_0^{\max\{2\theta-1, 1\}}$  and  $u \in [2, p_0]$ .

The remainder of the proof of Lemma 5.4.2 consists of a union bound over a  $1/2$ -net of a  $2k$ -dimensional unit sphere,  $\mathcal{C}(T, k)^2$  pairs of linear subspaces and an arbitrary set  $\mathcal{N}$ . Therefore, we can conclude with probability at least  $1 - \exp(-u + 2 \log \mathcal{C}(T, k) + \log(5)2k + \log |\mathcal{N}|)$ ,

$$\sup_{t, t_0 \in T, v \in \mathcal{N}} \frac{\left| \frac{1}{m} \epsilon_i y_i^v \langle a_i, t - t_0 \rangle \right|}{\|t - t_0\|_2} \leq 2eC_\theta K \sqrt{\frac{u}{m}},$$

if  $m \geq p_0^{\max\{2\theta-1, 1\}}$  and  $u \in [2, p_0]$ .

The relatively straightforward substitutions towards a simpler bound shows that to get non-trivial results, we will need to put additional restriction on the number of sub-Weibull moments  $p_0$ . Say that  $u' = u - 2 \log \mathcal{C}(T, k) - \log(5)2k - \log |\mathcal{N}|$ , such that with probability at least  $1 - e^{-u'}$  we have

$$\sup_{t, t_0 \in T, v \in \mathcal{N}} \frac{\left| \frac{1}{m} \epsilon_i y_i^v \langle a_i, t - t_0 \rangle \right|}{\|t - t_0\|_2} \leq 2eC_\theta K \sqrt{\frac{u' + 2 \log \mathcal{C}(T, k) + \log(5)2k + \log |\mathcal{N}|}{m}},$$

if  $m \geq p_0^{\max\{2\theta-1, 1\}}$  and

$$u' \in [2 - 2 \log \mathcal{C}(T, k) - \log(5)2k - \log |\mathcal{N}|, p_0 - 2 \log \mathcal{C}(T, k) - \log(5)2k - \log |\mathcal{N}|].$$

For non-trivial bounds on the probability we would like that  $u' > 0$  and this requires

$$p_0 > 2 \log \mathcal{C}(T, k) + \log(5)2k + \log |\mathcal{N}|, \quad (6.1)$$

thus the number of sub-Weibull moments required increases with the complexity of the signal set and will increase even further in later steps of the proof. For now, let us conclude this sub-section with the weak moment assumption version of Lemma 5.4.2.

**Lemma 6.2.5.** *Let  $\mathcal{N}$  be a finite set and define*

$$\text{Complexity} := 2 \log \mathcal{C}(T, k) + \log(5)2k + \log |\mathcal{N}|.$$

*If  $u > 0$  satisfies*

$$2 - \text{Complexity} \leq u \leq p_0 - \text{Complexity}, \quad \text{and}$$

$$m \geq \max \left\{ C_\theta \frac{\lambda^2 K^2}{\epsilon^2} (u + \text{Complexity}), p_0^{\max\{2\theta-1, 1\}} \right\}$$

*then with probability at least  $1 - e^{-u}$ ,*

$$\sup_{t, t_0 \in T, v \in \mathcal{N}} \frac{\left| \frac{1}{m} \epsilon_i y_i^v \langle a_i, t - t_0 \rangle \right|}{\|t - t_0\|_2} \leq \frac{\epsilon}{64\lambda},$$

*where  $C_\theta > 0$  is some constant dependent on  $\theta$ .*

## 6.2.4 Variance Bound - Hamming Distance

To obtain a uniform bound of the Hamming distance with high probability like in Lemma 5.4.3, we required two parts. The second part, Lemma 5.4.5, used the Chernoff Bound combined with a bound that does not depend on the distribution of the measurement vectors and noise. The first part, Lemma 5.4.4, relied on the distribution of the measurement vectors through the term

$$\sup_{z \in B_2^n} \mathbb{P}(|\langle a_1, z \rangle| \geq \eta/\delta).$$

We will now try to replace the sub-exponential tail bound with a weak moment based tail bound like in Lemma 6.2.2. If  $\theta = 1$ , i.e., the moments behave sub-exponential, then letting

$$\delta := \frac{\eta}{eK} \log(\lambda/\eta),$$

would result in

$$\sup_{z \in B_2^n} \mathbb{P}(|\langle a_1, z \rangle| \geq \eta/\delta) \leq \exp\left(-\frac{\eta}{eK\delta}\right) = \frac{\eta}{\lambda},$$

with the condition that  $\eta/\delta \in [2eK, eKp_0]$ .

The issue is however with

$$\eta := c_2(\lambda + K) \sqrt{\frac{\log \mathcal{C}(T, k) + u'}{m}},$$

as required for Lemma 5.4.5, we get that

$$\eta/\delta = eK \log(\eta/\lambda) \leq eK \log(1/4) < eK,$$

where the second to last inequality is because for Lemma 5.4.3 we require that  $\eta/\lambda \leq 1/4$ .

To fix this issue, we would like to have a concentration inequality that extends towards zero. For this, we can slightly weaken Lemma 6.2.2, by increasing the probability enough such that the inequality also holds on  $[0, 2]$ .

**Lemma 6.2.6.** *Let  $0 < 2 < p_0 \leq \infty$  and  $\theta > 0$ . If  $X$  is a random variable that satisfies*

$$\|X\|_{L^p} \leq Kp^\theta,$$

for some  $K > 0$  and for all  $p \in [2, p_0]$ , then

$$\mathbb{P}(|X| \geq eKu^\theta) \leq ce^{-u},$$

for all  $u \in [0, p_0]$ , for some  $c > 0$ .

It is also at this part of the proof where the set  $\mathcal{N}$  for Lemma 6.2.5 is chosen such that

$$\log |\mathcal{N}| \leq C \left( \log \mathcal{C}(T, k) + k \log(5R) + k \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right) \right),$$

such that for the lower bound on the number of moments from equation (6.1) it will be enough to hold that

$$p_0 > C \left( \log \mathcal{C}(T, k) + k \log(5R) + k \log \left( \frac{m}{\log \mathcal{C}(T, k) + u'} \right) \right),$$

or

$$p_0 > C (\log \mathcal{C}(T, k) + k \log(5R) + k \log(m)).$$

## 6.2.5 Variance Bound - Sign Differences

The final component to the variance bound is bounding the term

$$\sup_{P_1, P_2 \in \mathcal{P}, \bar{t} \in (P_1 - P_2) \cap B_2^n} \max_{|I| = \lfloor 4m\eta/\lambda \rfloor} \frac{2}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle|,$$

like in Lemma 5.4.7. This is where following the same steps as the proof in the previous chapter will result in a contradiction.

From the centralization inequality

$$\| |\langle a_i, t \rangle| - \mathbb{E} |\langle a_i, t \rangle| \|_{L^p} \leq 2 \| \langle a_i, t \rangle \|_{L^p}, \quad \text{for } p \geq 1,$$

it follows with probability at least  $1 - e^{-u_2}$  that

$$\frac{1}{|I|} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \leq 4eC_\theta K \sqrt{\frac{u_2}{|I|}}, \quad (6.2)$$

if  $|I| \geq p_0^{\max\{2\theta-1, 1\}}$  and  $u_2 \in [2, p_0]$ .

In the next step, we take a union bound over all possible subsets  $I \subset [m]$  that satisfy  $|I| = \lfloor 4m\eta/\lambda \rfloor$ . For simplicity, assume that  $4m\eta/\lambda$  is a natural number. The resulting lower bound on the probability then becomes

$$1 - \exp\left(-u_2 + \log\left(\binom{m}{4m\eta/\lambda}\right)\right).$$

To get non-trivial results, we would like that  $\log\left(\binom{m}{4m\eta/\lambda}\right) \leq u_2$ , which requires

$$\log\left(\binom{m}{4m\eta/\lambda}\right) \leq p_0,$$

because  $u_2$  cannot be larger than  $p_0$ . Now note that

$$4\frac{\eta}{\lambda} \log\left(\frac{\lambda}{4\eta}\right) \leq \log\left(\binom{m}{4m\eta/\lambda}\right) \leq 4\frac{\eta}{\lambda} \log\left(\frac{e\lambda}{4\eta}\right).$$

Compare this to the constraint

$$p_0 \leq |I| = \frac{4\eta}{\lambda}.$$

As discussed in the proof of Lemma 5.4.6,  $\frac{4\eta}{\lambda}$  is at most 1 due to the sampling complexity and becomes arbitrarily small when  $m$  becomes even larger. Hence, as  $m$  becomes larger,  $\log\left(\frac{\lambda}{4\eta}\right)$  becomes greater than one, implying that

$$\log\left(\binom{m}{4m\eta/\lambda}\right) = \frac{4\eta}{\lambda} + x \leq p_0 \leq \frac{4\eta}{\lambda},$$

for some  $x > 0$  cannot hold. Another way to see the problem is that if one writes out  $\log\left(\frac{e\lambda}{4\eta}\right)$ , one gets an additional  $\log(m)$  factor. Therefore, the lower bound on  $p_0$  increases asymptotically faster than the upper bound.

Instead of trying to directly apply concentration inequalities, we could also try to derive concentration properties of the smaller sum by concentration properties of the sum over all  $i \in [m]$ . While we can use Jensen's inequality to find

$$\begin{aligned} \left\| \sup_{t \in S} \frac{1}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \right\|_{L_p} &= \left\| \sup_{t \in S} \mathbb{E} \left[ \frac{1}{m} \sum_{i \in I} |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \middle| a_i \text{ for all } i \in I \right] \right\|_{L_p} \\ &\leq \left\| \sup_{t \in S} \frac{1}{m} \sum_{i=1}^m |\langle a_i, \bar{t} \rangle| - \mathbb{E} |\langle a_i, \bar{t} \rangle| \right\|_{L_p}, \end{aligned}$$

for some set  $S$ , we cannot also take the supremum over the different subsets  $I$  into consideration with this inequality. This is because the conditioning depends on the choice of  $I$ . Using the union bound on all subsets  $|I|$  after using this method only on all the vectors  $\bar{t}$  will result in the same contradiction as before.

Sadly, we will end this chapter without proving recovery guarantees with heavy tailed measurements and leave the final lemma unanswered. At the end of the next chapter we will observe through numerical experiments that the behaviour for some heavy tailed distributions is similar to the sub-Gaussian and sub-Exponential results. Furthermore, there exist results for other compressed sensing problems that make use of the weak moment assumption, see for example Lecué and Mendelson [24] or Dirksen et al. [10]. Hence, it is to be expected that this problem will someday be solved.



# Chapter 7

## Numerical experiments

In the previous chapters we have seen various theoretical guarantees on one-bit compressed sensing with generative models. In this chapter we will see various experimental results, focusing on the difference between reconstruction methods that exploit sparsity with respect to a basis and methods that use generative models. We will also look into the impact of dithering. Most experiments have been done using the MNIST data set, with some final experiments to show similar behaviour for the more complex CIFAR-10 data set.

### 7.1 Normalized scenario

Recall that in the normalized scenario as treated in Section 3.1, we consider measurements of the form

$$y_i = \text{sign}(\langle a_i, t_0 \rangle + \xi_i), \quad \text{for } i = 1, \dots, m.$$

Without control of the noise, we can only hope to reconstruct the signal up to normalization, hence we are interested in measuring the normalized error

$$\left\| \frac{t_0}{\|t_0\|_2} - \frac{t}{\|t\|_2} \right\|_2.$$

Alternatively, one could use the normalized geodesic distance

$$\frac{1}{\pi} \arccos \left( \left\langle \frac{t_0}{\|t_0\|_2}, \frac{t}{\|t\|_2} \right\rangle \right).$$

#### 7.1.1 Algorithms

Most theoretical guarantees for one-bit compressed sensing that were discussed in earlier chapters were in the form of minimizing

$$\min_{t \in T} L(t) := \|t\|_2^2 - \frac{2\lambda}{m} y^T A t,$$

for some  $\lambda$  and signal set  $T$ . Because this optimization problem has both a correlation and regularization term, we will refer to this optimization problem as the regularized problem.

This optimization problem can be solved numerically using a projected gradient descent method [16], resulting in an iterative method:

$$t_{k+1} = P_T(t_k - \delta \nabla L(t_k)), \quad \text{with} \quad (7.1)$$

$$\nabla L(t_k) = 2t_k - \frac{2\lambda}{m} A^T y,$$

where  $\delta$  is the learning rate and  $P_T(t)$  is the closest element in  $T$  to  $t$ , i.e., the generally non-linear projection of  $t$  onto  $T$ .

The projection  $P_T(t)$  can be efficiently computed when  $T$  is the set of  $s$ -sparse vectors  $\Sigma_s^n$  by setting all but the  $s$  absolute largest values to zero. Similar algorithms based on this efficient projection like Iterative Hard Thresholding have been successful in unquantized compressed sensing [1]. By transforming a signal into another basis, we can also relatively efficiently compute the projection onto the  $s$ -sparse vectors in any basis. However, when  $T$  is the range of a generative model  $G$ , the projection becomes an optimization problem without a simple solution.

PyTorch is used for the experiments in this chapter. This machine learning framework can train neural networks, but also train the input of a neural network instead of its parameters, therefore resulting in a simple, yet relatively expensive, implementation for solving the intermediate optimization problem

$$P_{G(X)}(t) := G(\arg \min_{x \in X} \|G(x) - t\|_2), \quad (7.2)$$

when the generative model  $G$  is a neural network. This algorithm is given in Listing 7.1. Note that instead of only using the decoder part of the auto-encoder, the whole auto-encoder is used. Practically, this makes no big difference, but it does show that we do not need to know about the low dimensional bottleneck in the network to be able to benefit from it. In some cases, the resulting signal set can also be smaller than when only using the decoder.

Listing 7.1: Projection algorithm

```
def projection(im):
    inp = im.clone()
    inp.requires_grad_(True)

    optimizer = torch.optim.Adam([inp], lr=learning_rate)

    for i in range(num_epochs):
        optimizer.zero_grad()
        out = autoEncoder(inp)
        loss = torch.norm(im-out)
        loss.backward()
        optimizer.step()

    return autoEncoder(inp).detach()
```

Another algorithm proposed by Jacques et al. [19, 18] for one-bit compressed sensing on  $\Sigma_s^n$  is Binary Iterative Hard-Thresholding (BIHT), which is defined as

$$t_{k+1} = P_{\Sigma_s^n} (t_k + \delta A^T (y - \text{sign}(At_k))),$$

which is a projected sub-gradient descent method for the optimization problem

$$\min_{t \in T} \| [y \odot At]_- \|_1,$$

with  $\odot$  being the Hadamard product and  $[x]_- := \min\{x, 0\}$ . This method penalizes only incorrect signs.

As proposed by Liu et al. [26], we can use the same algorithm with a projection onto the range of a generative model and we will refer to this algorithm as Binary Iterative Projection (BIP).

## 7.1.2 Experimental results

Unless stated otherwise, the following assumptions and notations will be used in the remainder of this chapter:

- Any error is the average of three separate reconstructions with different realizations of the random measurement matrix and noise.
- The measurement matrices  $A$  consist of element-wise i.i.d. distributed random variables with mean zero and variance one.
- Gaussian pre-quantization noise with mean zero and standard deviation 0.1 has been added.
- When using a generative model, the number of layers in the encoder and decoder is denoted by  $l$  and the bottleneck with  $k$ .
- The activation functions in the neural networks are ReLU for all layers except the last one, for which the sigmoid activation function is used.
- Each neural network has been training using an Adam optimizer with mean square error loss function for 20 epochs with a learning rate of 0.001.
- The projection on the range of a generative model is implemented as in Listing 7.1, with a learning rate of 0.001 and 50 iterations when the projection is used as part of a reconstruction algorithm and a learning rate of 0.0001 and 200 iterations when used to project the true image to the range of the generative model.

For precise details on the implementations, the code used for the experiments can be found at [github.com/jeverink/MastersThesis](https://github.com/jeverink/MastersThesis).

Both the regularized algorithm and binary iterative projection require the expensive operation of projecting a signal onto the range of the generative model, so the cost of a single iteration is approximately equal for both algorithms.



Figure 7.1: First five elements of the MNIST test set.

To test the two algorithms, we apply them to the reconstruction of the first five elements from the test set of MNIST [25] as seen in Figure 7.1.

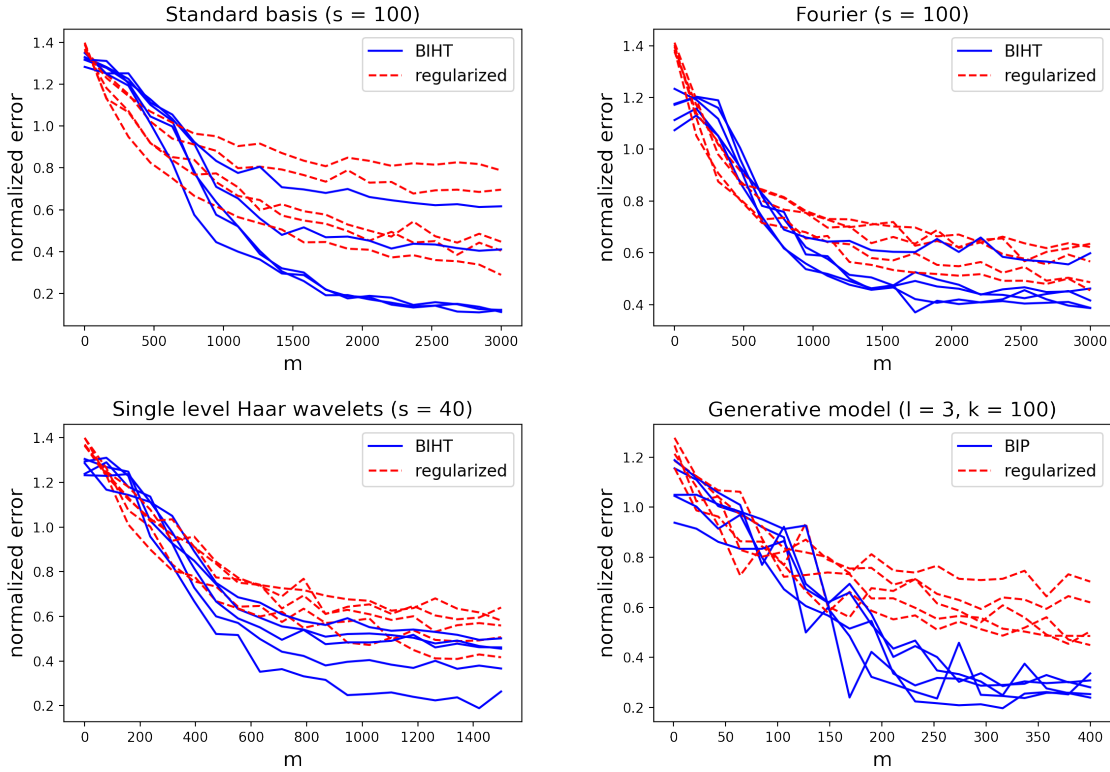
Figure 7.2 shows how the accuracy improves when increasing the number of measurements for both the regularized and binary iterative hard-thresholding/projection methods assuming sparsity in the standard basis, Fourier basis, single level Haar wavelets [5] and using a generative model. The choice of sparsity and bottleneck has been chosen empirically. For the behaviour for varying sparsity and bottleneck, see Figure 7.6.

First, note that the regularized methods initially decrease faster, yet when the binary iterative hard-thresholding/projection overtakes, it quickly reaches a better accuracy. Because the binary iterative hard-thresholding/projection, once quickly decreasing, reaches a better accuracy and it does not require tuning a regularization parameter, we will use this method for further experiments in the normalized scenario.

Second, the level of sparsity has quite an impact on the achievable accuracy for some images, hence the spread of the lines for larger  $m$ . One can improve the achievable accuracy by increasing the sparsity level, but this will also increase the number of measurements required for a good reconstruction. The generative model is trained to compress all the different classes of images, hence the achievable accuracy can generally be more uniform over all the images and therefore results in less spread of the lines for large  $m$ .

Third, the sparsity in the standard basis is clearly not as well a representation of the data set as using Haar wavelets, reaching similar if not better results with less than half the sparsity and almost half the measurements. The generative model requires even less measurements, but seems to result in less consistent results.

Although the normalized errors allows for good quantitative comparison of the methods, visually comparing the images allows us to qualitatively compare the difference between the sparsity and generative model assumptions. Figures 7.3 and 7.4 show reconstructions with increasing  $m$  compared to the directly projecting the true signal onto the signal set using sparsity and a generative model.



algorithm parameters			
		Learning rate	iterations
Standard, Fourier and Haar	BIHT	0.0005	500
	regularized	0.1	250
Generative model	BIP	0.02	50
	regularized	0.8	50

Figure 7.2: Normalized error for the reconstruction of the first five images from the test set of MNIST, each image a line, using both the regularized and binary iterative hard-thresholding/projection algorithms for varying number of measurements.

Standard basis ( $s=100$ )

$m = 300$	$m = 500$	$m = 700$	$m = 1000$	$m = 1500$	$m = 2000$	$m = 2500$	projector
acc 1.234	acc 1.132	acc 0.890	acc 0.608	acc 0.277	acc 0.175	acc 0.189	acc 0.049
acc 1.189	acc 1.138	acc 1.050	acc 0.629	acc 0.549	acc 0.421	acc 0.416	acc 0.307
acc 1.129	acc 1.078	acc 0.663	acc 0.490	acc 0.257	acc 0.169	acc 0.151	acc 0.000
acc 1.283	acc 1.200	acc 1.069	acc 0.828	acc 0.756	acc 0.722	acc 0.649	acc 0.516
acc 1.177	acc 1.049	acc 0.935	acc 0.560	acc 0.293	acc 0.293	acc 0.112	acc 0.059

Single level Haar wavelets ( $s=40$ )

$m = 200$	$m = 300$	$m = 400$	$m = 500$	$m = 600$	$m = 700$	$m = 800$	projector
acc 1.017	acc 0.949	acc 0.580	acc 0.554	acc 0.556	acc 0.508	acc 0.434	acc 0.246
acc 1.154	acc 0.970	acc 0.938	acc 0.607	acc 0.527	acc 0.493	acc 0.543	acc 0.337
acc 1.021	acc 0.666	acc 0.647	acc 0.620	acc 0.491	acc 0.418	acc 0.315	acc 0.143
acc 1.224	acc 1.041	acc 0.675	acc 0.679	acc 0.578	acc 0.567	acc 0.473	acc 0.340
acc 1.211	acc 1.102	acc 0.934	acc 0.679	acc 0.657	acc 0.684	acc 0.618	acc 0.364

Figure 7.3: Reconstructions using Binary Iterative Hard Thresholding for varying number of measurements and assuming sparsity within the standard basis and Haar wavelets. A learning rate of 0.0005 for 500 iterations has been used.

Generative model ( $l=3, k=50$ )



















































$m = 10$	$m = 50$	$m = 100$	$m = 125$	$m = 150$	$m = 175$	$m = 200$	$m = 250$	$m = 300$	projector
									
acc 1.182	acc 1.087	acc 0.892	acc 0.456	acc 0.606	acc 0.261	acc 0.273	acc 0.254	acc 0.248	acc 0.188
									
acc 1.020	acc 0.935	acc 0.845	acc 0.746	acc 0.492	acc 0.513	acc 0.624	acc 0.413	acc 0.411	acc 0.243
									
acc 1.174	acc 1.225	acc 0.423	acc 0.591	acc 0.415	acc 0.293	acc 0.190	acc 0.194	acc 0.214	acc 0.135
									
acc 0.916	acc 0.882	acc 0.678	acc 0.608	acc 0.391	acc 0.232	acc 0.371	acc 0.226	acc 0.273	acc 0.168
									
acc 1.076	acc 0.908	acc 1.050	acc 1.041	acc 0.459	acc 0.461	acc 0.397	acc 0.311	acc 0.361	acc 0.234

Figure 7.4: Reconstructions using Binary Iterative Projection using a generative model. A learning rate of 0.02 for 50 iterations has been used.

Perhaps the biggest difference between the sparsity in the standard basis or Haar wavelets in Figure 7.3 and the generative model in Figure 7.4 is that in the former the noise decreases while the true image increases, while in the latter the image shift from various blobs towards the real image. This difference in behaviour results in a problem when visually assessing whether a solution is true or at least believable. In the sparsity images of Figure 7.3 one can visually judge that the solution consists of mostly noise, while in the generative model images of Figure 7.4 one can get solutions that might be misleading. Thus one has to be very careful when assessing solutions obtained from very little measurements through a generative model, because such a model contains much less extraneous signals.

Having seen that a generative model can greatly reduce the number of measurements needed to obtain accurate solutions, we have yet to see how the complexity of the model influences the accuracy and sampling complexity. Because of the steep descent seen for the Binary Iterative Hard-Thresholding/Projection in Figure 7.2, we will measure how many measurements a signal set requires by considering for what value of  $m$  this steep descent happens. For this, we will call the threshold  $m$  the first value of  $m$  for which accuracy passes halfway between the approximately worst accuracy obtained by reconstructing with one measurement and the accuracy obtained by projecting the true signal onto the signal set. How this threshold  $m$  is determined is

visualized in Figure 7.5. How this value behaves compared to the error of directly projecting onto the range of the generative model and the sparsity level and bottleneck size is shown in Figure 7.6.

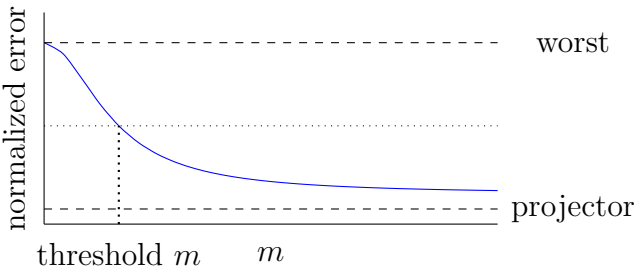
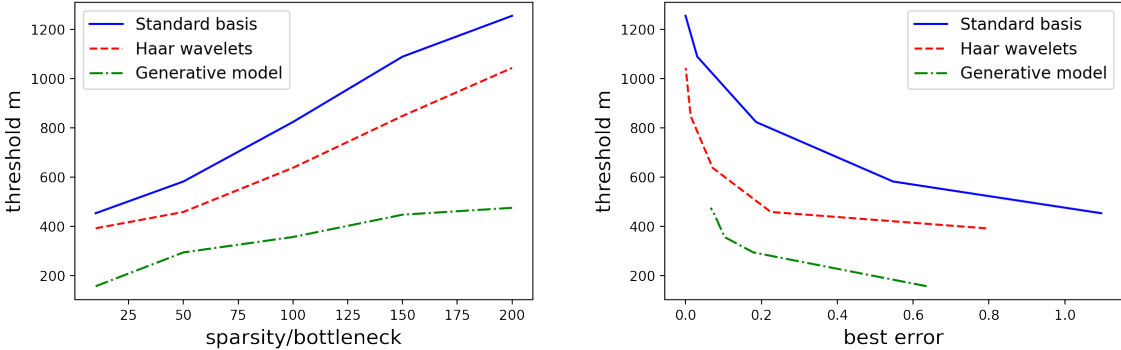


Figure 7.5: Visualization of the threshold  $m$ , the approximate number of measurements needed for a reasonable reconstruction.



algorithm parameters		
	Learning rate	iterations
Standard basis	0.001	500
Haar wavelets	0.001	500
Generative model	0.001	75

Figure 7.6: Approximate number of measurements needed for a reasonable reconstruction using different models when varying the sparsity/bottleneck. Averaged over five images. The left plot shows the impact of the sparsity, while the right plot shows how the same data compares to the approximate best possible error as obtained by directly projecting the true image on the signal set. The encoder and decoder each consist of 1 layer.

In Figure 7.6 we can see that great reduction in required measurements as seen earlier holds for many networks. Only when you want to greatly increase the accuracy by increasing the complexity of the network or the bottleneck does the number of measurements required greatly increase, however, as seen in Figures 7.3 and 7.4 the



additional details obtained in this final stretch are far from necessary to speak of a good solution in this specific data set.

### 7.1.3 Union of smaller generative models

Instead of using a single, large neural network as generative model for the MNIST data set, we can also use multiple smaller neural networks, one for each label. The natural interpretation is that the data set is not sampled from one large manifold, but that each number has its own smaller manifold. Hence, instead of working with the signal set  $G(X)$ , we will consider  $\cup_{i=0}^9 G_i(X)$ . The hope is that by splitting into multiple generative models, the whole signal set becomes simpler by not containing signals to connect the different labels.

You could replace the projection on the larger generative model by multiple projections and then choose the best one, however, this can cause issues when the smaller generative models are well separated. Instead, we can reconstruct the signal in each generative model separately and then choose the reconstruction with the best objective value. In some sense, we are actually combining the reconstruction algorithm with a sample-based classification algorithm. Figure 7.7 shows the results of such an algorithm. We can see that when the number of measurements is very low, multiple labels can give optimal objective values and therefore, no decision can be made. Yet, when increasing the number of measurements, the objective value of the reconstruction from the true label will increase the slowest, thus for enough measurements, the smaller, true generative model will dominate.

Multi model reconstruction of "1"

m										
10										
20										
30										
50										
100										
projector										

Figure 7.7: Reconstructions using multiple, label-specific generative models. The ones marked with a red border in each row are the ones with the lowest objective value, hence could be chosen as final solution. The final row shows the images obtained by directly projecting the true image on the range of each generative model. The encoder and decoder part of each generative model consists of 2 layers with a bottleneck of 20. A learning rate of 0.02 for 50 iterations was used.

## 7.2 Dithering

Recall that if we also want to reconstruct the norm of the signal, we consider measurements of the form

$$y_i = \text{sign}(\langle a_i, t_0 \rangle + \xi_i + \tau_i), \text{ for } i = 1, \dots, m,$$

where the  $\tau_i$  are i.i.d. uniformly distributed on  $[-\lambda, \lambda]$  for some  $\lambda > 0$ . The regularized optimization problem as seen in previous chapters can be solved using the same gradient projection method as seen in Equations (7.1), where the  $\lambda$  used for dithering and regularization are the same.

## 7.2.1 Sparsity versus generative model

Before we can start comparing the different structural assumptions, we need to make a choice for the dithering parameter  $\lambda$ . Assuming that there is no noise, then any dithering beyond the radius of the signal set is useless. As we will be using only a little Gaussian noise, we will use  $\lambda = 18$  for the following numerical experiment, which is slightly larger than the radius of the MNIST data set.

Figure 7.8 shows how the error decays when increasing the number of measurements for various structural assumptions. It also shows the normalized error and the error of the norm. Similar to the normalized scenario, the generative model gives the best results, then the Haar wavelets and finally the standard basis. It looks like the error of the norm seems to stagnate or even slightly increase while mostly the normalized error seems to continue improving.

## 7.2.2 Optimal regularization

In the previous experiments, the choice of the dithering constant  $\lambda$  has been 18, which is approximately the radius of the MNIST data set. However, while this value results in enough dithering to distinguish between all the elements in the signal set that could reasonably be correct, this does not have to be optimal. From the results in the previous chapters, we know that increasing the regularization parameter should increase the number of required measurements, but how does it behave for relatively low  $\lambda$ ?

Figures 7.9 and 7.10 show for both sparsity in the standard basis and using a generative model the error when reconstructing using various  $\lambda$ . It is important to notice that choosing  $\lambda$  too large is just as bad, if not worse, in terms of the error as choosing  $\lambda$  far too low, thus choosing a good regularization parameter can drastically impact the number of measurements required for good reconstructions. We also see that the optimal regularization parameter differs with the signal that is being reconstructed. The optimal choice of  $\lambda$  seems to correlate with how well we can reconstruct that specific signal, as is expected from the lower bounds on  $\lambda$  found in earlier chapters, thus choosing optimal  $\lambda$  a priori without additional information about the signal is out of the question.

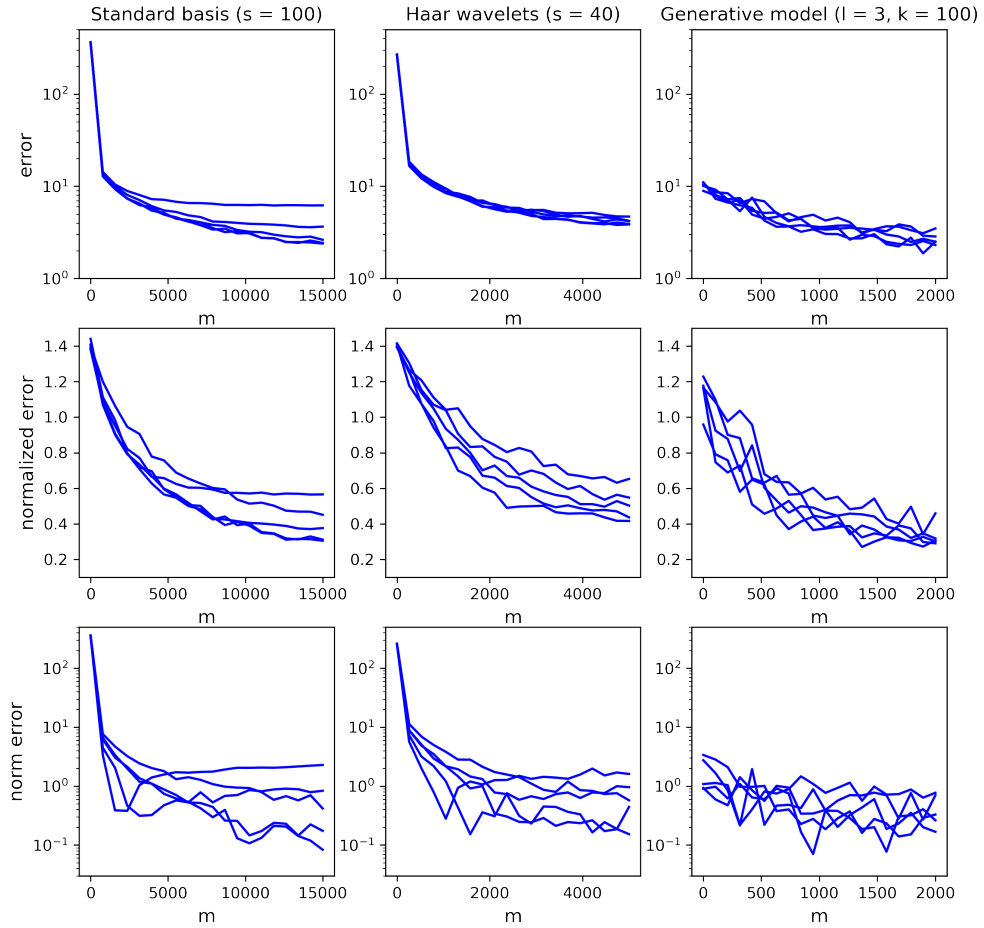


Figure 7.8: Errors when reconstructing the first five elements from the MNIST data set with dithering ( $\lambda = 18$ ). A learning rate of 1 for 100 iterations was used for the standard basis and Haar wavelets, while a learning rate of 0.2 for 50 iterations was used for the Generative model.

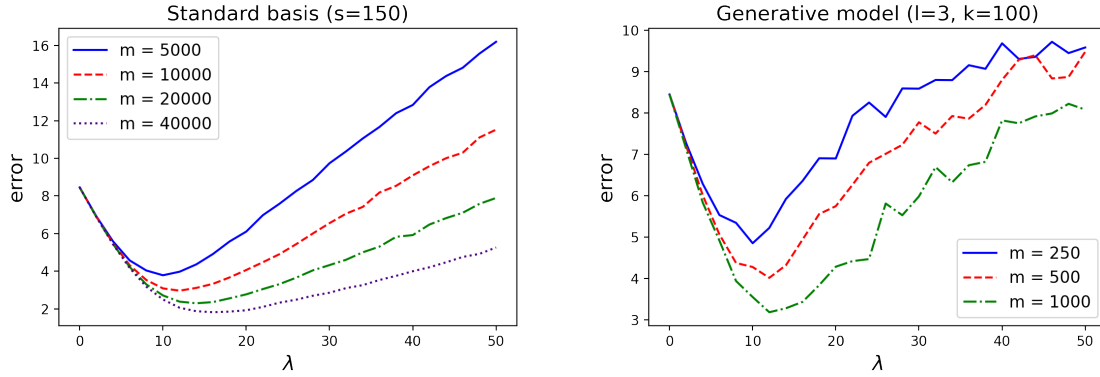


Figure 7.9: Average error of reconstructing the first five elements from the CIFAR-10 test set with varying regularization parameter  $\lambda$  and number of measurements  $m$ . A learning rate of 0.1 for 100 iterations was used for the standard basis and learning rate of 0.1 for 100 iterations was used for the Generative model.

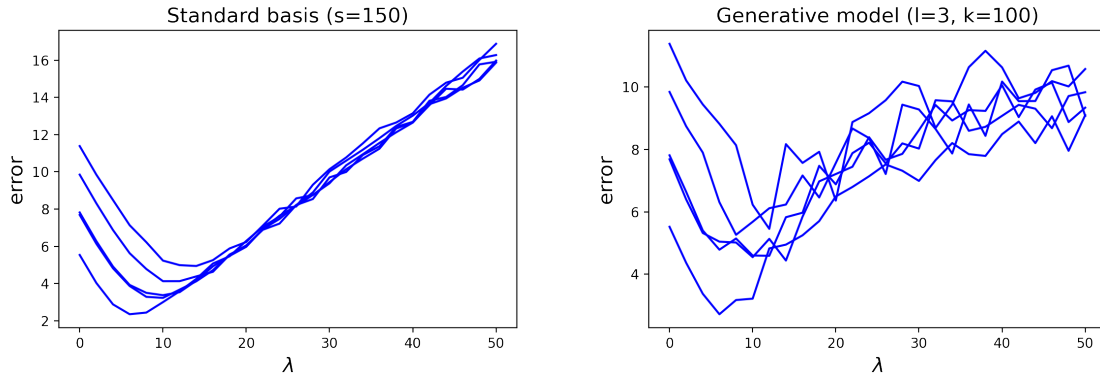


Figure 7.10: Errors for different test images with 5000 measurements for various regularization parameter  $\lambda$ . A learning rate of 0.1 for 100 iterations was used for the standard basis and learning rate of 0.1 for 100 iterations was used for the Generative model.

It turns out that the optimal choice of the dithering parameter  $\lambda$  correlates with norm of the signal. For the first 3216 images in the MNIST test set, the approximate optimal choice of  $\lambda$  is plotted against their norm in Figure 7.11. In the sparsity scenario it clearly shows a correlation between optimal  $\lambda$  and norm, and that the optimal lambda increases with the number of measurements. In the generative model scenario, even though the algorithm for finding the optimal value of  $\lambda$  has been greatly affected by the inconsistent recovery as seen in Figure 7.10, shows similar correlation.

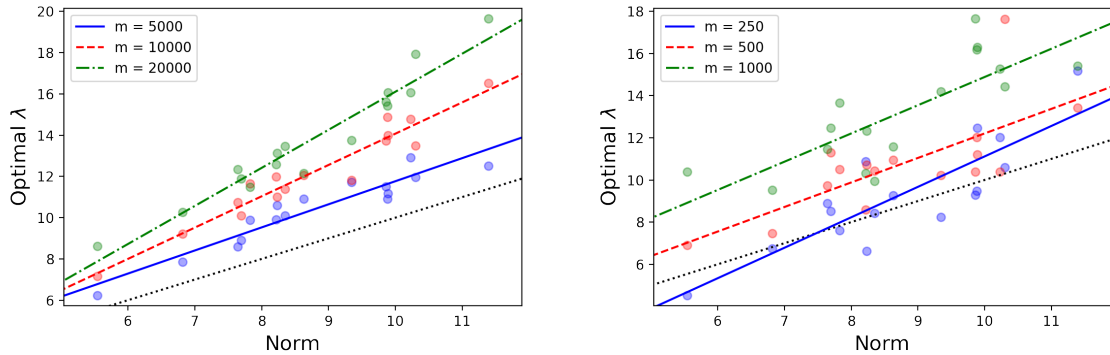


Figure 7.11: Approximation of the optimal  $\lambda$  for the first 16 images of the MNIST test set in terms of the norm of the image, together with a linear fitted line. A learning rate of 0.05 for 200 iterations was used for standard sparsity and a learning rate of 0.2 for 50 iterations was used for the Generative model. The optimal value of  $\lambda$  is approximated using ternary search.

### 7.2.3 A more complex data set: CIFAR-10

We have only looked at the relatively simple MNIST data set as of yet. Hence we will now finish these numerical experiments by looking at the more complex CIFAR-10 data set [22], which consists of colored images with  $32^2$  pixels. A sample of the CIFAR-10 data set is given in Figure 7.12.

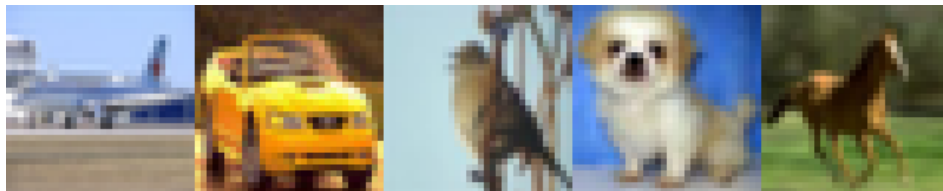


Figure 7.12: Sample of CIFAR-10 data set.

As basis for sparsity, we consider single level Haar wavelets for each color channel of the image. Figure 7.13 shows that behaviour for increasing  $m$  for this sparsity assumption and a generative model is similar to the behaviour on the MNIST data set.

### 7.2.4 Stronger noise

Though a very little bit of noise has been added to the previous experiments, we have yet to look at the impact of stronger noise. Figure 7.14 shows what happens when we increase the variance of the Gaussian pre-quantization noise. As expected from the theoretical results, stronger noise requires more measurements and stronger dithering. Perhaps surprising is what happens when over-dithering. The experimental results show that when choosing  $\lambda$  too large, the noise has no impact on the error. Or worded

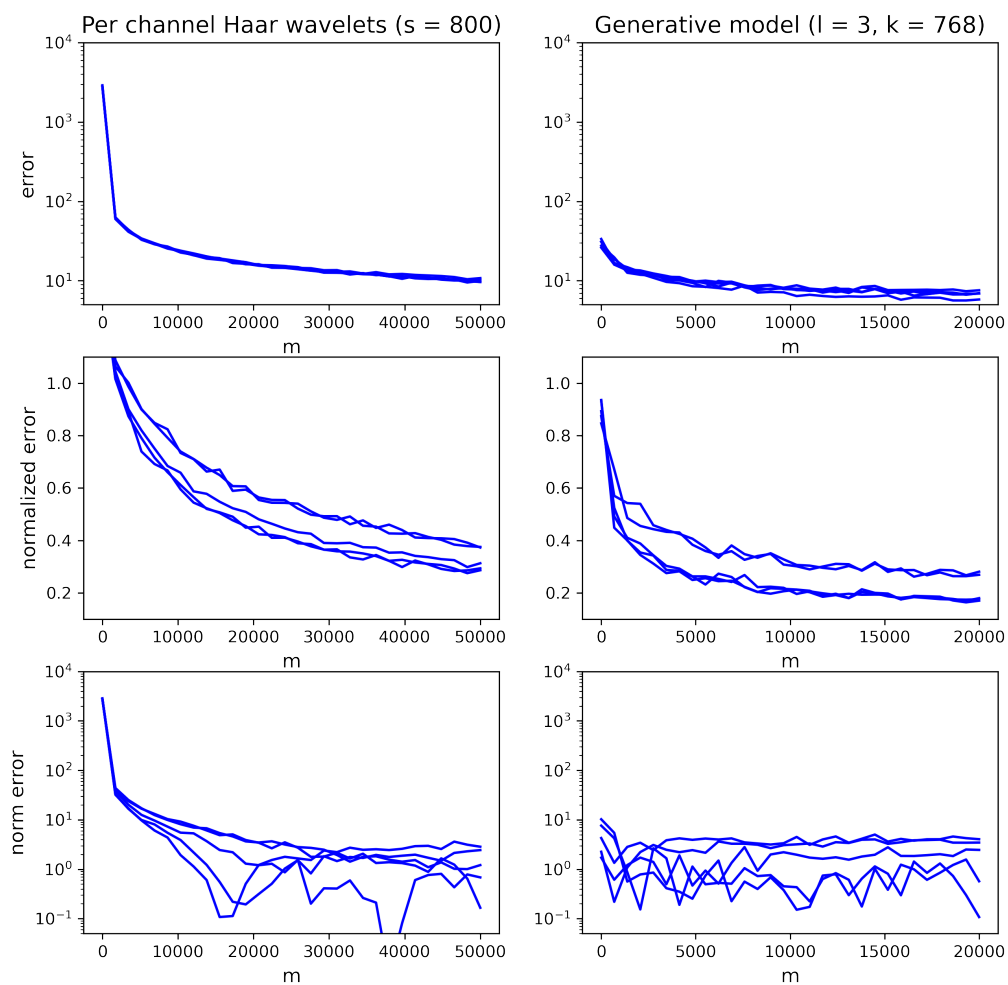


Figure 7.13: Errors of reconstructing the first five images from the CIFAR-10 data set with dithering ( $\lambda = 60$ ). A learning rate of 0.01 for 200 iterations was used for the Haar wavelets and a learning rate of 0.2 for 50 iterations for the generative model.

differently, over-dithering gives results similar to stronger noise. It is likely that as the dithering becomes the dominant noise term, the outer values of the dithering are more detrimental than helpful.

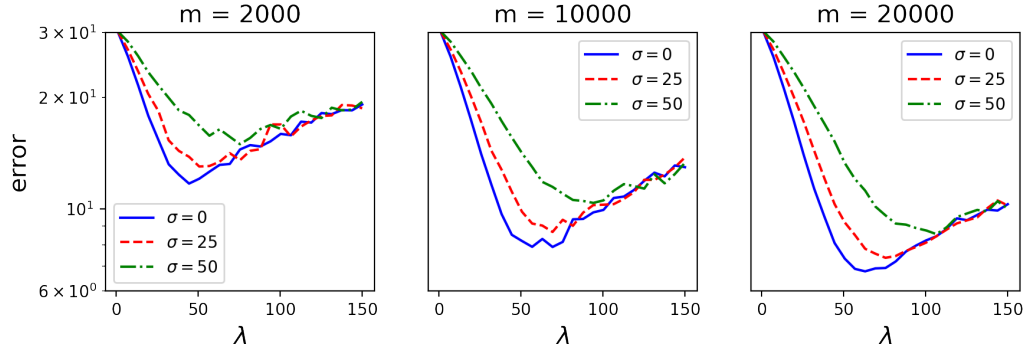


Figure 7.14: Average error of reconstructing five images using a generative model with various noise levels. A learning rate of 0.2 for 50 iterations was used.

### 7.2.5 Heavier tailed measurements

In Chapter 5, we have seen that sub-Gaussian and sub-Exponential measurement vectors should give similar results and in Chapter 6 we were close to showing that even the heavy tailed Student's-t distribution should behave similar. So as a final experiment, let us look at these different measurement vectors. Figure 7.15 shows the error for varying  $\lambda$ ,  $m$  and distributions. The data is nearly indistinguishable and, perhaps surprisingly, even the Student's-t distribution with only 3 degrees-of-freedom, which does not have a third moment, behaves sub-Gaussian.

## 7.3 Which assumption is better?

When looking at the experimental results, they would suggest that using a generative model is the more efficient choice. However, we are not even close to exhausting all possible choices of basis, generative models and data sets.

From an algorithmic point of view, sparsity has the benefit of being easily structured and therefore having a relatively easy constraint to work with. In contrast, the geometry of the range of a generative model can be very complex and therefore has an expensive projection, but it has the benefit of using modern machine learning methods to learn this geometry from a data set.

Learning the geometry can sometimes be relatively easy, like in the case of natural data sets like MNIST and CIFAR-10, as seen in the previous experiments. However, not all data sets are easy to learn. To take it to the extreme, consider the signal set of sparse vectors  $\Sigma_{100}^{784}$ , which we used in the MNIST experiments earlier. The MNIST



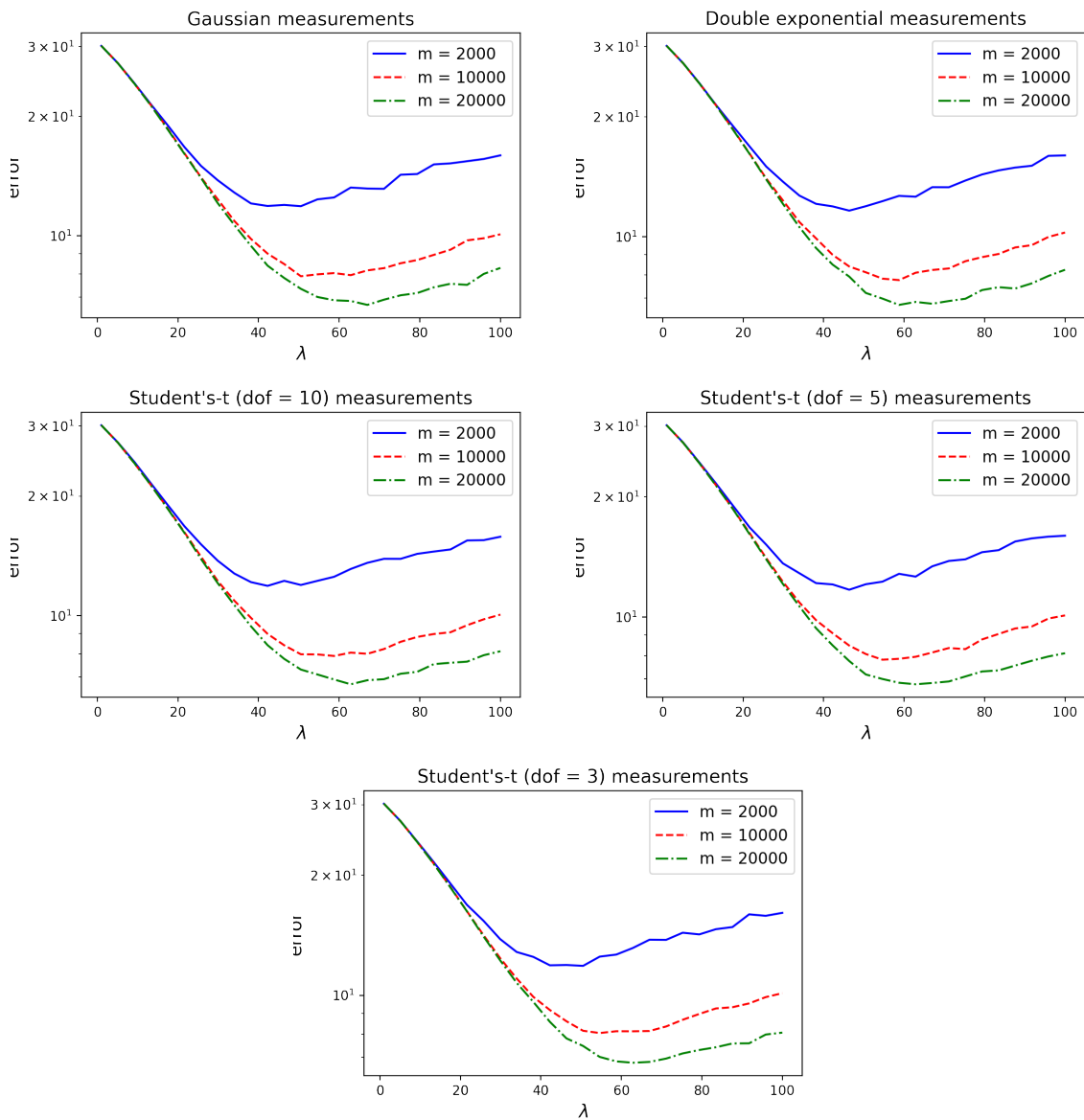


Figure 7.15: The average error over five images, when reconstructing from measurements with different standardized measurement vector distributions. A learning rate of 0.2 for 25 iterations was used.

data set would only occupy a small part of this set, but the whole set satisfies

$$\mathcal{C}_{\text{lin}}(\Sigma_{100}^{784}, 100) = \binom{784}{100} \approx 10^{128}.$$

Training a neural network with relatively small bottleneck to model the many subspaces of  $\Sigma_{100}^{784}$  is not an easy task, while working with  $\Sigma_{100}^{784}$  is straightforward. The complexity of the resulting generative model will probably result in requiring more measurements than when using sparsity.

Whether to work with sparsity or a generative model can therefore be highly dependent on the application. We cannot conclude from these limited experiments that the use of a generative model is better for natural data sets in general.

## Chapter 8

### Conclusion and further research

In this thesis we have seen how using a generative model, instead of assuming traditional sparsity, can be very powerful in one-bit compressed sensing. We have shown that the number of measurements required for reconstruction depends linearly on the radius of the signal set and the dimension of the latent space, while it only depends logarithmically on the complexity of the generative model in terms of the Lipschitz constant or the number of linear pieces.

We have seen that the statistical recovery guarantee on sub-exponential measurement vectors and noise by Qiu et al. [30] for one-bit compressed sensing with dithering can be generalized to signal sets that consist of piecewise, low-dimensional linear pieces, which includes sparsity in a basis and piecewise linear neural networks. The noise can be further generalized to only require more than the first two moments at the cost of stronger dithering and therefore requires more measurements. It is highly likely that the measurement vectors can also be generalized to only require a few well-behaved moments, however the arguments used caused a contradiction in the final step.

Even though the numerical experiments were conducted with relatively simple data sets and neural networks, the reduction in the required number of measurements was substantial, although at the cost of computation time and possible issues with the visual assessment of the reconstructions. Using dithering to also reconstruct the magnitude of the signal is observed to be a powerful technique, but if the dithering is not well-tuned then it can even worsen the reconstruction.

Even though the experiments show that using a generative model in one-bit compressed sensing is very powerful, one has to decide whether the reduced number of measurements is worth the great increase in required computational power.

### Further research

The incompleteness of the argument for extending from the sub-exponential setting to the heavy-tailed setting is perhaps the most frustrating part of this thesis. It is only in the final lemma where the argumentation used causes a contradiction, hence finding a fix for this part of the argument or trying to find an alternative route towards using heavy-tailed distributions is an important topic of further study.

In the (sub-)Gaussian part of this thesis we have seen that Lipschitz continuous

functions are a natural choice for generative models, yet in the sub-exponential and heavy-tailed part we only considered piecewise linear generative models. Although both can be used with the Gaussian width in the sub-Gaussian setting, the piecewise linearity is quite a limitation in the heavier tailed settings. Hence, an important topic for further research is to further generalize the set of permissible signal sets and generative models in the sub-exponential and heavy-tailed settings. In the best case, another quantity similar to Gaussian width could work with heavier tailed distributions.

In this thesis, we only considered independent and identically distributed measurement vectors. Such measurements are not very realistic and an important topic of research within compressed sensing is the study of various kinds of structured matrices, for example, Fourier sub-sampling and partial circulant matrices.

The numerical experiments conducted were limited to the relatively simple MNIST and CIFAR-10 data sets. Furthermore, the neural networks used were relatively simple and not perfectly trained. To truly test whether generative models are beneficial in practical scenarios, tests could be carried out with more complex and realistic data sets, for example, those used in medical imaging, together with well trained deep neural networks.

The projection onto the range of a generative model is quite a bit more time expensive than projecting on the set of sparse vectors. Hence, whether the reduction of the number of measurements is worth the additional computation time of the methods considered in this thesis is application dependent. Other methods could probably be more efficient, for example, by considering the generative model as a part of the objective functions instead of a constraint or using deep learning to speed up the projection process. Therefore, further research could be done on improving the reconstruction algorithm with generative models.

# Appendix A

## Miscellaneous proofs

### Proof of lemma 2.4.3

Let  $G : X \subseteq \mathbb{R}^k \rightarrow \mathbb{R}^n$  be  $\gamma$ -Lipschitz. Choose some  $x_0 \in X$  and define the star-shaped version of  $G$  around  $G(x_0)$ ,  $G^* : X \times [0, 1] \rightarrow \mathbb{R}^n$  as

$$G^*(x, \lambda) := \lambda G(x) + (1 - \lambda)G(x_0).$$

Then the image  $G^*(T, [0, 1])$  is star-shaped around the chosen point  $G(x_0)$  and  $G^*$  is Lipschitz continuous with Lipschitz constant at most  $\sqrt{2}\gamma(1 + \text{diam}(X))$ .

*Proof.* That  $G^*(X, [0, 1])$  is star-shaped around  $G(x_0)$  follows directly from the definition of  $G^*$ .

For the Lipschitz constant, we start by splitting up the distance between any two points  $G^*(x, \lambda)$  and  $G^*(y, \mu)$  with  $(x, \lambda), (y, \mu) \in X \times [0, 1]$  as follows

$$\begin{aligned} \|G^*(x, \lambda) - G^*(y, \mu)\|_2 &\leq \|G^*(x, \lambda) - G^*(x, \frac{1}{2}(\lambda + \mu))\|_2 \\ &\quad + \|G^*(x, \frac{1}{2}(\lambda + \mu)) - G^*(y, \frac{1}{2}(\lambda + \mu))\|_2 \\ &\quad + \|G^*(y, \frac{1}{2}(\lambda + \mu)) - G^*(y, \mu)\|_2. \end{aligned}$$

For the middle term we get

$$\begin{aligned} \|G^*(x, \frac{1}{2}(\lambda + \mu)) - G^*(y, \frac{1}{2}(\lambda + \mu))\|_2 &\leq \frac{1}{2}(\lambda + \mu)\|G(x) - G(y)\|_2 \\ &\leq \gamma\|x - y\|_2. \end{aligned}$$

For the first term we get

$$\begin{aligned} \|G^*(x, \lambda) - G^*(x, \frac{1}{2}(\lambda + \mu))\|_2 &\leq \|\frac{1}{2}(\lambda - \mu)G(x) - \frac{1}{2}(\lambda - \mu)G(x_0)\|_2 \\ &\leq \frac{1}{2}|\lambda - \mu|\|G(x) - G(x_0)\|_2 \\ &\leq \frac{1}{2}\text{diam}(G(X))|\lambda - \mu|. \end{aligned}$$

Furthermore, we have that  $\text{diam}(G(X)) \leq \gamma \text{diam}(X)$ . Combining this with the same bound for the third term results in

$$\begin{aligned} \|G^*(x, \lambda) - G^*(y, \mu)\|_2 &\leq \gamma \|x - y\|_2 + \text{diam}(G(X)) |\lambda - \mu| \\ &\leq \gamma(1 + \text{diam}(X)) (\|x - y\|_2 + |\lambda - \mu|). \end{aligned}$$

Now using that for any  $a, b \in \mathbb{R}$  it holds that  $(a + b)^2 \leq 2(a^2 + b^2)$ , results in

$$\sqrt{(\|x - y\|_2 + |\lambda - \mu|)^2} \leq \sqrt{2} \sqrt{(\|x - y\|_2^2 + |\lambda - \mu|^2)} = \sqrt{2} \|(x, \lambda) - (y, \mu)\|_2,$$

thus we can conclude that

$$\|G^*(x, \lambda) - G^*(y, \mu)\|_2 \leq \sqrt{2} \gamma (1 + \text{diam}(X)) \|(x, \lambda) - (y, \mu)\|_2,$$

for any  $x, y \in X$  and  $\lambda, \mu \in [0, 1]$ , hence  $G^*$  is  $\sqrt{2} \gamma (1 + \text{diam}(X))$ -Lipschitz. ■

## Proof of lemma 4.1.2

If  $T \subseteq \mathbb{R}^n$ , then for any  $R \geq 0$ ,

$$w(T \cap RB_2^n) \leq CR \left( \sqrt{\log \mathcal{C}_{\text{lin/aff}}(T, k)} + \sqrt{k} \right),$$

for some constant  $C > 0$ .

*Proof.* Note that we only have to proof the lemma for the affine covering number, because  $\mathcal{C}_{\text{aff}}(T, k) \leq \mathcal{C}_{\text{lin}}(T, k)$ , and for simplicity we will abbreviate  $\mathcal{C}_{\text{aff}}(T, k)$  to  $\mathcal{C}$ .

Let  $p_i + V_i$  for  $i = 1, \dots, \mathcal{C}$  be  $k$ -dimensional affine sub spaces of  $\mathbb{R}^n$ , such that

$$T \subseteq \cup_{i=1}^{\mathcal{C}} p_i + V_i.$$

By the monotonicity of Gaussian width we find

$$w(T \cap RB_2^n) \leq w\left(\left(\cup_{i=1}^{\mathcal{C}} p_i + V_i\right) \cap RB_2^n\right).$$

Without loss of generality we can assume that  $p_i \in RB_2^n$ , because if we cannot find such  $p_i$ , then  $p_i + V_i$  and  $RB_2^n$  are disjoint. We can split the  $p_i$  from the  $V_i$  using

$$\begin{aligned} w\left(\left(\cup_{i=1}^{\mathcal{C}} p_i + V_i\right) \cap RB_2^n\right) &= \mathbb{E} \sup_{i \in [\mathcal{C}]} \sup_{v_i \in V_i: p_i + v_i \in RB_2^n} \langle g, p_i + v_i \rangle \\ &\leq \mathbb{E} \sup_{i \in [\mathcal{C}]} \sup_{v_i \in V_i: t_i + v_i \in p_i + 2RB_2^n} \langle g, p_i + v_i \rangle \\ &= \mathbb{E} \sup_{i \in [\mathcal{C}]} \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, p_i + v_i \rangle \\ &\leq \mathbb{E} \sup_{i \in [\mathcal{C}]} \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle + \mathbb{E} \sup_{i \in [\mathcal{C}]} \langle g, p_i \rangle \\ &= w\left(\left(\cup_{i=1}^{\mathcal{C}} V_i\right) \cap 2RB_2^n\right) + w(\{p_1, \dots, p_{\mathcal{C}}\}). \end{aligned}$$

By assumption and the Gaussian width of a finite set,  $w(\{p_1, \dots, p_C\}) \leq cR\sqrt{\log C}$  for some constant  $c > 0$ . Using a centralization argument results in

$$\begin{aligned} \mathbb{E} \sup_{i \in [C]} \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle &\leq \mathbb{E} \sup_{i \in [C]} \left| \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle \right| \\ &\leq \mathbb{E} \sup_{i \in [C]} \left| \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle - \mathbb{E} \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle \right| + \sup_{i \in [C]} \mathbb{E} \sup_{v_i \in V_i \cap 2RB_2^n} |\langle g, v_i \rangle|. \end{aligned}$$

For the second term we find that

$$\mathbb{E} \sup_{v_i \in V_i \cap 2RB_2^n} |\langle g, v_i \rangle| = w(V_i \cap 2RB_2^n) \leq 2R\sqrt{k}.$$

Next, by a concentration inequality for supremum of Gaussian processes (see e.g. Theorem 5.8 in [3]) we get that

$$\sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle - \mathbb{E} \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle,$$

is a centered  $2R$ -sub-Gaussian random variable, hence

$$\mathbb{E} \sup_{i \in [C]} \left| \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle - \mathbb{E} \sup_{v_i \in V_i \cap 2RB_2^n} \langle g, v_i \rangle \right| \leq cR\sqrt{\log C},$$

for some constant  $c > 0$ .

Combining all these bounds allows us to conclude that

$$w(T \cap RB_2^n) \leq CR \left( \sqrt{\log C} + \sqrt{k} \right),$$

for some absolute constant  $C > 0$ . ■

## Proof of decomposition of sign

**Lemma A.0.1.** *If  $Z$  and  $V$  are arbitrary random variables and  $\tau \sim \text{Unif}[-\lambda, \lambda]$ , then*

$$\mathbb{E}[Z \text{sign}(V + \tau) | Z, V] = \frac{ZV}{\lambda} \mathbf{1}_{\{|V| \leq \lambda\}} + Z \mathbf{1}_{\{V > \lambda\}} - Z \mathbf{1}_{\{V < -\lambda\}}.$$

*Proof.* First note that

$$\mathbb{E}[Z \text{sign}(V + \tau) | Z, V] = \mathbb{E}[Z \text{sign}(V + \tau) \mathbf{1}_{\{|V| \leq \lambda\}} | Z, V] + Z \mathbf{1}_{\{V > \lambda\}} - Z \mathbf{1}_{\{V < -\lambda\}},$$

i.e., two cases where the sign is deterministic, independent of the value of  $\tau$  and one term where it actually depends on  $\tau$ .

As for the first term in the summation, this is equivalent to

$$\begin{aligned}\mathbb{E}[Z\text{sign}(V + \tau)\mathbf{1}_{\{|V|\leq\lambda\}}|Z, V] &= Z (\mathbb{P}(V + \tau \geq 0) - \mathbb{P}(V + \tau \leq 0)) \mathbf{1}_{\{|V|\leq\lambda\}} \\ &= Z \left( \frac{\lambda + V}{2\lambda} - \frac{\lambda - V}{2\lambda} \right) \mathbf{1}_{\{|V|\leq\lambda\}} \\ &= Z \frac{V}{\lambda} \mathbf{1}_{\{|V|\leq\lambda\}},\end{aligned}$$

concluding the proof. ■

## Tail bound - Bias

**Lemma A.0.2.** *For any random variable  $V$  and  $\lambda > 0$  we have*

$$\|V\mathbf{1}_{\{|V|>\lambda\}}\|_{L_2}^2 \leq \lambda^2\mathbb{P}(|V| > \lambda) + 2 \int_{\lambda}^{\infty} t\mathbb{P}(|V| > t) dt.$$

*Proof.* For any random variable  $X$  we have

$$\mathbb{E}|X|^2 = \int_0^{\infty} \mathbb{P}(|X|^2 > t) dt = 2 \int_0^{\infty} t\mathbb{P}(|X| > t) dt.$$

Now letting  $X = V\mathbf{1}_{\{|V|>\lambda\}}$  and splitting the integral results in

$$\|V\mathbf{1}_{\{|V|>\lambda\}}\|_{L_2}^2 = 2 \int_0^{\lambda} t\mathbb{P}(|V|\mathbf{1}_{\{|V|>\lambda\}} > t) dt + 2 \int_{\lambda}^{\infty} t\mathbb{P}(|V|\mathbf{1}_{\{|V|>\lambda\}} > t) dt.$$

For  $t \in (0, \lambda)$ ,  $\mathbb{P}(|V|\mathbf{1}_{\{|V|>\lambda\}} > t)$  is constant as  $|V|\mathbf{1}_{\{|V|>\lambda\}}$  almost surely takes the value 0 or a value higher than  $\lambda$ , so the first term reduces to

$$2\mathbb{P}(|V|\mathbf{1}_{\{|V|>\lambda\}} > \lambda) \int_0^{\lambda} t dt = \lambda^2\mathbb{P}(|V|\mathbf{1}_{\{|V|>\lambda\}} > \lambda).$$

For  $t \geq \lambda$ , it does not matter that  $\mathbf{1}_{\{|V|>\lambda\}}$  reduces any value of  $|V|$  below  $\lambda$  to 0, hence  $\mathbb{P}(|V|\mathbf{1}_{\{|V|>\lambda\}} > t) = \mathbb{P}(|V| > t)$ . Therefore, the first term further simplifies to

$$\lambda^2\mathbb{P}(|V| > \lambda),$$

while the second term simplifies to

$$2 \int_{\lambda}^{\infty} t\mathbb{P}(|V| > t) dt,$$

completing the proof. ■



## Proof Lemma 5.4.3

Let  $\eta > 0$  and  $\epsilon \in [0, 1/2]$ . If with probability at least  $1 - p$ ,

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| \geq \eta\}} \leq \epsilon,$$

and with probability at least  $1 - q$ ,

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| < \eta\}} \leq \epsilon,$$

then with probability at least  $1 - p - q$ ,

$$\sup_{t_0 \in T} d_H(y^{t_0}, y^{\mathcal{N}(t_0)}) \leq 2\epsilon.$$

*Proof.* By a union bound it holds with probability at least  $1 - p - q$  that both

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| \geq \eta\}} \leq \epsilon,$$

and

$$\sup_{t_0 \in T} \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| < \eta\}} \leq \epsilon,$$

hold.

The key to the proof is the observation that if  $\mu$  is a probability measure or a normalized counting measure, and  $A$  and  $B$  are two events that both satisfy  $\mu(A), \mu(B) \geq 1/2 + \epsilon$  with  $\epsilon \in [0, 1/2]$ , then from the equality  $\mu(A \cap B) + \mu(A \cup B) = \mu(A) + \mu(B)$  it follows that  $\mu(A \cap B) \geq 2\epsilon$ .

Now fix arbitrary  $t_0 \in T$  and let  $\mu$  be a normalized counting measure over  $[m]$  with  $A$  the set consisting of the  $i$  for which  $|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| < \eta$  and  $B$  the set consisting of the  $i$  for which  $|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| \geq \eta$ , such that  $\mu(A) \geq 1 - \epsilon$  and  $\mu(B) \geq 1 - \epsilon$  by assumption. Therefore  $\mu(A \cap B) \geq 1 - 2\epsilon$ , hence at least  $1 - 2\epsilon$  percent of the  $i$  satisfy both  $|\langle a_i, t_0 - \mathcal{N}(t_0) \rangle| < \eta$  and  $|\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i| \geq \eta$ , and therefore  $\text{sign}(\langle a_i, t_0 \rangle + \xi_i + \tau_i) = \text{sign}(\langle a_i, \mathcal{N}(t_0) \rangle + \xi_i + \tau_i)$ .

From this we can conclude that

$$d_H(y^{t_0}, y^{\mathcal{N}(t_0)}) \leq 2\epsilon,$$

for any  $t_0 \in T$  with probability at least  $1 - p - q$ , concluding the proof. ■

## Missing step in the proof of Lemma 4.2.1

For simplicity we will denote  $\log \mathcal{C}(T, k)$  and  $k$  by  $x$  and  $y$  respectively.

**Lemma A.0.3.** *There exists constant  $c > 0$  such that for any  $x, y > 0$ ,*

$$2x + 4y \log \left( \frac{ecx}{2y} \right) < cx.$$

.

*Proof.* We can rewrite the inequality we want to proof as

$$\frac{4}{c-2} \log \left( \frac{ec}{2} \right) + \frac{4}{c-2} \log(z) < z,$$

with  $z := x/y$  and assuming that  $c > 2$ .

The first term that depends only on  $c$  converges to 0 as  $c$  goes to infinity, thus for  $c$  large enough, it will be enough to prove that

$$0.5 + 0.5 \log(z) < z.$$

This follows from the fact that the function  $f(z) = 0.5 + \log(z) - z$  is concave with maximum at  $z = 0.5$  with value  $f(z) = \log(0.5) < 0$ . ■

# Appendix B

## High-dimensional probability theory

Because concepts from high-dimensional probability theory come up throughout this whole thesis, this appendix contains a small overview of some of the definitions and theorems used. The main focus is on the definitions and some important properties of sub-Gaussian and sub-Exponential random variables/vectors and concentrations of their sums. Most of this theory can be found in extended form in the book High-Dimensional Probability: An Introduction with Applications in Data Science by Roman Vershynin [32]. Some small proofs not found in this book are provided.

### B.1 Sub-Gaussian random variables

Random variables that have similar behaviour to Gaussian random variables satisfy some of the same strong concentration phenomena, hence the following definition characterizes this similar behaviour.

**Definition B.1.1.** (Sub-Gaussian random variable) A random variable  $X$  is called **sub-Gaussian** if any of the following equivalent definitions holds:

1. There exists a constant  $C_1$  such that

$$\mathbb{P}(|X| \geq u) \leq 2e^{-u^2/C_1^2} \quad \text{for all } u \geq 0.$$

2. There exists a constant  $C_2$  such that

$$\|X\|_{L_p} \leq C_2\sqrt{p} \quad \text{for all } p \geq 1.$$

3. There exists a constant  $C_3$  such that

$$\mathbb{E} \exp(X^2/C_3^2) \leq 2.$$

The **sub-Gaussian norm** is defined by

$$\|X\|_{\psi_2} := \inf\{K > 0 : \mathbb{E} \exp(X^2/K^2) \leq 2\},$$

and satisfies

$$\mathbb{P}(|X| \geq u) \leq 2e^{-cu^2/\|X\|_{\psi_2}^2} \quad \text{for all } u \geq 0,$$

for some  $c > 0$ .

Two standard examples of sub-Gaussian random variables are Gaussian random variables and bounded random variables.

Although we will not directly use the concentration of sums of sub-Gaussian random variables in this thesis, we will see a form of Hoeffding's inequality for comparison with the sub-exponential Bernstein's inequality in the next section.

**Theorem B.1.2** (Hoeffding's inequality, Theorem 2.6.2 in [32]). *Let  $X_1, \dots, X_m$  be i.i.d. mean zero, sub-Gaussian random variables, then for any  $u \geq 0$ , we have*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m X_i\right| \geq c\|X_1\|_{\psi_2}\sqrt{\frac{u}{m}}\right) \leq 2e^{-u},$$

for some universal constant  $c > 0$ .

## B.2 Sub-exponential random variables

Weakening the quadratic behaviour of sub-Gaussian random variables to linear behaviour result in the following definition.

**Definition B.2.1.** (Sub-Exponential random variable) A random variable  $X$  is called **sub-exponential** if any of the following equivalent definition holds:

1. There exists a constant  $C_1$  such that

$$\mathbb{P}(|X| \geq u) \leq 2e^{-u/C_1} \quad \text{for all } u \geq 0.$$

2. There exists a constant  $C_2$  such that

$$\|X\|_{L_p} \leq C_2 p \quad \text{for all } p \geq 1.$$

3. There exists a constant  $C_3$  such that

$$\mathbb{E} \exp(|X|/C_3) \leq 2.$$

The **sub-exponential norm** is defined by

$$\|X\|_{\psi_1} := \inf\{K > 0 : \mathbb{E} \exp(|X|/K) \leq 2\},$$

and satisfies

$$\mathbb{P}(|X| \geq u) \leq 2e^{-cu/\|X\|_{\psi_1}} \quad \text{for all } u \geq 0,$$

for some  $c > 0$ .

Standard examples of sub-exponential variables includes Laplace random variables, bounded random variables and all sub-Gaussian random variables.

A fundamental concentration inequality that we will use in this thesis is Bernstein's inequality for the sum of sub-exponential random variables.

**Theorem B.2.2** (Bernstein's inequality). *Let  $X_1, \dots, X_m$  be i.i.d. mean zero, sub-exponential random variables, then for any  $u \geq 0$ , we have*

$$\mathbb{P} \left( \left| \frac{1}{m} \sum_{i=1}^m X_i \right| \geq c \|X_1\|_{\psi_1} \left( \sqrt{\frac{u}{m}} + \frac{u}{m} \right) \right) \leq 2e^{-u},$$

for some universal constant  $c > 0$ .

*Proof.* From Corollary 2.8.3 in [32] we get that under the assumptions of the theorem that for any  $t \geq 0$ ,

$$\mathbb{P} \left( \left| \frac{1}{m} \sum_{i=1}^m X_i \right| \geq t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{\|X_1\|_{\psi_1}^2}, \frac{t}{\|X_1\|_{\psi_1}} \right) m \right).$$

Now let  $t = K \|X_1\|_{\psi_1} \left( \sqrt{\frac{u}{m}} + \frac{u}{m} \right)$  for some  $K > 0$ , then we find that

$$\frac{t^2}{\|X_1\|_{\psi_1}^2} m = K^2 u + 2K^2 u \sqrt{\frac{u}{m}} + K^2 \frac{u^2}{m} \geq K^2 u,$$

and

$$\frac{t}{\|X_1\|_{\psi_1}} m = K (\sqrt{um} + u) \geq Ku.$$

Hence,

$$2 \exp \left( -c \min \left( \frac{t^2}{\|X_1\|_{\psi_1}^2}, \frac{t}{\|X_1\|_{\psi_1}} \right) m \right) \leq 2 \exp (-cu \min(K^2, K)).$$

Now choose  $K$  large enough such that  $\min(K^2, K) \geq 1/c$  finishes the proof. ■

## B.3 Random vectors

To extend the definition of sub-Gaussian and sub-exponential random variables to random vectors, consider the following definition.

**Definition B.3.1.** (Sub-Gaussian and sub-exponential random vectors) A random vector  $X$  is called **sub-Gaussian** if for any  $x \in \mathbb{R}^n$ ,  $\langle X, x \rangle$  is sub-Gaussian. We define the **sub-Gaussian norm for random vectors** by

$$\|X\|_{\psi_2} := \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_2}.$$

Similarly, a random vector  $X$  is called **sub-exponential** if for any  $x \in \mathbb{R}^n$ ,  $\langle X, x \rangle$  is sub-Gaussian. We define the **sub-Gaussian norm for random vectors** by

$$\|X\|_{\psi_1} := \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_1}.$$

Note that we use the name notation for the sub-Gaussian/sub-exponential norm for random variables and random vectors. Which one is being used should be clear in the context.

As an example, any random vector with i.i.d. mean zero, sub-Gaussian or sub-exponential random variables is a sub-Gaussian or sub-exponential random vector respectively.

We do not want measurement vectors  $a$  that are biased towards specific directions. Hence, if we consider the marginal  $\langle a, x \rangle$  for unit vector  $x$ , we would like this value to be constant for any unit vector  $x$ . The following definition

**Definition B.3.2.** (Isotropic random vector) A random vector  $X$  in  $\mathbb{R}^n$  is called **isotropic** if

$$\mathbb{E}\langle X, x \rangle \langle X, y \rangle = \langle x, y \rangle,$$

for all  $x, y \in \mathbb{R}^n$ .

For a mean zero random vector, being isotropic is equivalent to the covariance matrix being the unit matrix.

**Lemma B.3.3.** A random vector  $X$  in  $\mathbb{R}^n$  is isotropic if and only if

$$\mathbb{E}XX^T = I_n,$$

where  $I_n$  the  $n \times n$  identity matrix.

*Proof.* Assume that  $\mathbb{E}XX^T = I_n$ , then for any  $x, y \in \mathbb{R}^n$  we have

$$\mathbb{E}\langle X, x \rangle \langle X, y \rangle = \mathbb{E}x^T X X^T y = x^T \mathbb{E}X X^T y = x^T I_n y = \langle x, y \rangle.$$

If  $X$  is isotropic, then for any  $x, y \in \mathbb{R}^n$  we have

$$x^T y = \mathbb{E}\langle X, x \rangle^2 = x^T \mathbb{E}X X^T y.$$

Thus, for any  $i, j \in [n]$  we have

$$[\mathbb{E}X X^T]_{i,j} = e_i^T \mathbb{E}X X^T e_j = e_i^T e_j = \delta_{ij},$$

hence  $\mathbb{E}X X^T = I_n$ . ■

From Lemma B.3.3 it directly follows that random vectors i.i.d. elements with mean zero and unit variance are isotropic.

## B.4 Chernoff bound

A final important concentration inequality for the sum of Bernoulli random variables is the Chernoff bound. Because we will use a different version than used in the High-dimensional probability book [32], a proof will be given similar to that in the book.

**Theorem B.4.1** (Chernoff bound). *Let  $X_1, \dots, X_m$  be i.i.d. Bernoulli random variables with parameter  $\mu$ , i.e.,  $\mathbb{E}X_i = \mathbb{P}(X_i = 1) = \mu = 1 - \mathbb{P}(X_i = 0)$ , then for  $\alpha > 0$  we have*

$$\mathbb{P}\left(\frac{1}{m} \sum_{i=1}^m X_i \geq (1 + \alpha)\mu\right) \leq \exp\left(\frac{-\alpha^2 m \mu}{2 + \alpha}\right).$$

*Proof.* For  $\lambda > 0$  and  $\alpha > 0$  we have

$$\mathbb{P}\left(\sum_{i=1}^m X_i \geq (1 + \alpha)m\mu\right) \leq e^{-\lambda(1+\alpha)m\mu} \prod_{i=1}^m \mathbb{E}e^{\lambda X_i},$$

The moment generating function of a Bernoulli random variable  $X_i$  with parameter  $\mu$  satisfies

$$\mathbb{E}e^{\lambda X_i} = e^\lambda \mu + (1 - \mu)1 + (e^\lambda - 1)\mu \leq \exp((e^\lambda - 1)\mu),$$

hence,

$$\mathbb{P}\left(\sum_{i=1}^m X_i \geq (1 + \alpha)m\mu\right) \leq [\exp(-\lambda(1 + \alpha) + (e^\lambda - 1))]^{m\mu}.$$

Now choose  $\lambda = \log(1 + \alpha)$  to conclude that

$$\mathbb{P}\left(\sum_{i=1}^m X_i \geq (1 + \alpha)m\mu\right) \leq \exp((\alpha - (1 + \alpha) \ln(1 + \alpha)) m \mu) \leq \exp\left(\frac{-\alpha^2 m \mu}{2 + \alpha}\right).$$

■

# Bibliography

- [1] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.
- [2] A. Bora, A. Jalal, E. Price, and A. G. Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017.
- [3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- [4] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE, 2008.
- [5] S. L. Brunton and J. N. Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2019.
- [6] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.
- [7] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.
- [8] S. Dirksen. Quantized compressed sensing: a survey. In *Compressed Sensing and Its Applications*, pages 67–95. Springer, 2019.
- [9] S. Dirksen, M. Iwen, S. Krause-Solberg, and J. Maly. Robust one-bit compressed sensing with manifold data. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5. IEEE, 2019.
- [10] S. Dirksen, G. Lécué, and H. Rauhut. On the gap between restricted isometry properties and sparse recovery conditions. *IEEE Transactions on Information Theory*, 64(8):5478–5487, 2016.
- [11] S. Dirksen, J. Maly, and H. Rauhut. Covariance estimation under one-bit quantization. *arXiv preprint arXiv:2104.01280*, 2021.
- [12] S. Dirksen and S. Mendelson. Non-gaussian hyperplane tessellations and robust one-bit compressed sensing. *arXiv preprint arXiv:1805.09409*, 2018.
- [13] S. Dirksen and S. Mendelson. Robust one-bit compressed sensing with partial circulant matrices. *arXiv preprint arXiv:1812.06719*, 2018.
- [14] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. 2013.
- [15] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [16] O. Güler. *Foundations of optimization*, volume 258. Springer Science & Business Media, 2010.



- [17] M. A. Iwen, F. Krahmer, S. Krause-Solberg, and J. Maly. On recovery guarantees for one-bit compressed sensing on manifolds. *Discrete & Computational Geometry*, pages 1–46, 2021.
- [18] L. Jacques, K. Degraux, and C. D. Vleeschouwer. Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing, 2013.
- [19] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors, 2015.
- [20] K. Knudson, R. Saab, and R. Ward. One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory*, 62(5):2748–2758, 2016.
- [21] S. Kotz and S. Nadarajah. *Multivariate t-distributions and their applications*. Cambridge university press, 2004.
- [22] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [23] A. K. Kuchibhotla and A. Chakraborty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.
- [24] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions, 2015.
- [25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [26] Z. Liu, S. Gomes, A. Tiwari, and J. Scarlett. Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors. In *International Conference on Machine Learning*, pages 6216–6225. PMLR, 2020.
- [27] Z. Liu and J. Scarlett. Information-theoretic lower bounds for compressive sensing with generative models. *IEEE Journal on Selected Areas in Information Theory*, 1(1):292–303, 2020.
- [28] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [29] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.
- [30] S. Qiu, X. Wei, and Z. Yang. Robust one-bit recovery via relu generative networks: Near optimal statistical rate and global landscape analysis. *arXiv preprint arXiv:1908.05368*, 2019.
- [31] A. W. Van Der Vaart and J. A. Wellner. Weak convergence and empirical processes. Springer, 1996.
- [32] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [33] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.
- [34] M. M. Wolf. *Mathematical foundations of supervised learning*, 2018.