

# A Guide to Probably Approximately Correct Bounds for Neural Networks

Thomas Walker

---

## CONTENTS

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>PAC</b>	<b>4</b>
2.1	Introducing PAC Bounds . . . . .	4
2.1.1	Notation . . . . .	4
2.1.2	PAC Bounds . . . . .	4
2.1.3	Occam Bounds . . . . .	5
2.2	Expected Risk Minimization . . . . .	5
2.3	Compression . . . . .	6
2.3.1	Establishing the Notion of Compression . . . . .	6
2.3.2	Compression of a Linear Classifier . . . . .	7
2.3.3	Compression of a Fully Connected Network . . . . .	9
<b>3</b>	<b>Empirical PAC-Bayes Bounds</b>	<b>12</b>
3.1	Introduction to PAC-Bayes Theory . . . . .	12
3.1.1	Bayesian Machine Learning . . . . .	12
3.1.2	Introducing PAC-Bayes Bounds . . . . .	12
3.1.3	PAC-Bayes Bounds . . . . .	13
3.2	Optimizing PAC-Bayes Bounds via SGD . . . . .	15
<b>4</b>	<b>Oracle PAC-Bayes Bounds</b>	<b>19</b>
4.1	Theory of Oracle PAC-Bayes Bounds . . . . .	19
4.1.1	Oracle PAC-Bayes Bounds in Expectation . . . . .	19

---

**AFFILIATION** Imperial College London

**CORRESPONDENCE** [thomas.walker21@imperial.ac.uk](mailto:thomas.walker21@imperial.ac.uk)

**DATE** July 2023

4.1.2	Oracle PAC-Bayes Bounds in Probability . . . . .	19
4.1.3	Bernstein's Assumption . . . . .	19
4.2	Data Driven PAC-Bayes Bounds . . . . .	20
4.2.1	Implementing Data-Dependent Priors . . . . .	20
<b>5</b>	<b>Extensions of PAC-Bayes Bounds</b>	<b>23</b>
5.1	Disintegrated PAC-Bayes Bounds . . . . .	23
5.1.1	Application to Neural Network Classifiers . . . . .	23
5.2	PAC-Bayes Compression Bounds . . . . .	24
5.3	Mutual Information . . . . .	25

## 1 INTRODUCTION

A great resource for introducing the field of PAC theory is given in [12]. It details the progression of results in the field and motivates the various research avenues. PAC learning theory is a more general field than just neural networks. However, my aim is to illustrate how this theory is being contextualized within machine learning, with a specific focus on neural networks. In [12] many of the applications of PAC theory were stated with accompanying references for the reader's interests. With this report, I want to elaborate on some of those points and provide further applications of the theory. This report will not provide an exhaustive summary of the theory of PAC bounds being applied to neural networks. I will provide some well-known results in the literature and additional results necessary for the applications we will discuss. It is recommended that for a comprehensive introduction to the field of PAC, the reader refers to [12]. Nevertheless, this report will be mostly self-contained, with proofs for the major results and elaboration on the specific implementations of PAC theory.

## 2 PAC

### 2.1 INTRODUCING PAC BOUNDS

**2.1.1 NOTATION** We will first introduce some basic notation that is for the most part consistent with [12] and will remain constant throughout the report. Along the way, we will need to introduce some more specialized notation for the different sections. The problems we will encounter will mainly be supervised classification tasks. Informally this means, we have a feature space  $\mathcal{X}$  and a label space  $\mathcal{Y}$  which combine to form the data space  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  for which some unknown  $\mathcal{D}$  is defined on. The challenge now is to learn a classifier  $h : \mathcal{X} \rightarrow \mathcal{Y}$  that correctly labels samples from  $\mathcal{X}$  according to  $\mathcal{D}$ . Formally,  $\mathcal{X}$  and  $\mathcal{Y}$  come associated with  $\sigma$ -algebras  $\mathcal{S}_x$  and  $\mathcal{S}_y$  respectively. Then  $\mathcal{D}$  is a probability distribution defined on the space  $(\mathcal{X} \times \mathcal{Y}, \mathcal{S}_x \otimes \mathcal{S}_y)$ . The training data  $S_m = \{(x_i, y_i)\}_{i=1}^m$  consists of  $m$  i.i.d samples from  $\mathcal{D}$ . As we are considering neural networks, a classifier will be parameterised by a weight vector  $\mathbf{w}$  which we will denote  $h_{\mathbf{w}}$ . Let  $\mathcal{W}$  denote the set of possible weights for a classifier and the set of all possible classifiers  $\mathcal{H}$  will sometimes be referred to as the hypothesis set. To assess the quality of a classifier we define a measurable function  $l : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$  called the loss function and we will assume that  $0 \leq l \leq C$ . As our training data is just a sample from the underlying (unknown) distribution  $\mathcal{D}$  there is the possibility that our classifier performs well on the training data, but performs poorly on the true distribution. Let the risk of our classifier be defined as

$$R(h_{\mathbf{w}}) = \mathbb{E}_{(x,y) \sim \mathcal{D}} (l(h(x), y)).$$

As our classifier is parameterised  $\mathbf{w}$  we will instead just write  $R(\mathbf{w})$  for the risk of our classifier. Similarly, we define the empirical risk of our classifier to be

$$\hat{R}(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m l(h_{\mathbf{w}}(x_i), y_i).$$

Note that  $\mathbb{E}_{S \sim \mathcal{D}^m} (\hat{R}(\mathbf{w})) = R(\mathbf{w})$ .

**2.1.2 PAC BOUNDS** Probably Approximately Correct bounds refer to a general class of bounds on the performance of a learned classifier. They aim to determine with high probability what the performance of a classifier will be like on the distribution  $\mathcal{D}$  when trained on some training data taken from this distribution.

**Theorem 2.1 ([12])** *Let  $|\mathcal{W}| = M < \infty$ ,  $\delta \in (0, 1)$ , and  $\mathbf{w} \in \mathcal{W}$  then it follows that*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( R(\mathbf{w}) \leq \hat{R}(\mathbf{w}) + C \sqrt{\frac{\log \left( \frac{M}{\delta} \right)}{2n}} \right) \geq 1 - \delta.$$

Theorem 2.1 says that with arbitrarily high probability we can bound the performance of our trained classifier on the unknown distribution  $\mathcal{D}$ . However, there is nothing to guarantee that the bound is

useful in practice. Note that requiring bounds to hold for greater precision involves sending  $\epsilon$  to 0 which increases the bound. If the bound exceeds  $C$  then it is no longer useful as we know already that  $R(\mathbf{w}) \leq C$ . It is important to note at this stage that there are two ways in which PAC bounds can hold. One set of bounds holds in expectation whilst the other hold in probability. Risk is a concept that will develop bounds in expectation. In Section 2.3 we will introduce definitions that will let us work with bounds that hold in probability. There are two general forms of PAC bounds, we have uniform convergence bounds and algorithmic-dependent bounds [11]. Uniform convergence bounds have the general form

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \sup_{\mathbf{w} \in \mathcal{W}} |R(\mathbf{w}) - \hat{R}(\mathbf{w})| \leq \epsilon \left( \frac{1}{\delta}, \frac{1}{m}, \mathcal{W} \right) \right) \geq 1 - \delta.$$

This can be considered as a worst-case analysis of hypothesis generalization, and so in practice will lead to non-vacuous bounds. Algorithmic-dependent bounds involve the details of a learning algorithm  $A$  and take the form

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( |R(A(S)) - \hat{R}(A(S))| \leq \epsilon \left( \frac{1}{\delta}, \frac{1}{m}, A \right) \right) \geq 1 - \delta.$$

These bounds can be seen as a refinement of the uniform convergence bounds as they are only required to hold for the output of the learning algorithm. It will be the subject of Section 5.1 to explore such bounds further.

**2.1.3 OCCAM BOUNDS** Occam bounds are derived under the assumption that  $\mathcal{H}$  is countable and that we have some bias  $\pi$  defined on the hypothesis space. Note that in our setup this does not necessarily mean that  $\mathcal{W}$  is countable, as multiple weights may correspond to the same classifier. However, as the Occam bounds hold true for all  $h \in \mathcal{H}$  it must also be the case that they hold for all classifiers corresponding to the weight  $\mathbf{w} \in \mathcal{W}$ . Using this we will instead assume that  $\pi$  is defined over  $\mathcal{W}$ .

**Theorem 2.2 ([5])** *Simultaneously for all  $\mathbf{w} \in \mathcal{W}$  and  $\delta \in (0, 1)$  the following holds,*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( R(\mathbf{w}) \leq \inf_{\lambda > \frac{1}{2}} \frac{1}{1 - \frac{1}{2\lambda}} \left( \hat{R}(\mathbf{w}) + \frac{\lambda C}{m} \left( \log \left( \frac{1}{\pi(\mathbf{w})} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \right) \geq 1 - \delta.$$

## 2.2 EXPECTED RISK MINIMIZATION

In light of Theorem 2.1 it may seem reasonable to want to identify the parameter value  $\hat{\mathbf{w}}_{\text{ERM}}$  that minimizes  $\hat{R}(\cdot)$ . This optimization process is known as Empirical Risk Minimization (ERM) and is formally defined as

$$\hat{\mathbf{w}}_{\text{ERM}} = \inf_{\mathbf{w} \in \mathcal{W}} \hat{R}(\mathbf{w}).$$

## 2.3 COMPRESSION

We now show how PAC bounds can be used to bound the performance of a compressed neural network. In classical statistical theory only as many parameters as training samples are required to overfit. So in practice, neural networks would be able to overfit the training data as they have many more parameters than training samples. Although overfitting to the training sample will yield a low empirical risk, in practice neural networks do not overfit to the data. This suggests that there is some capacity of the network that is redundant in expressing the learned function. In [7] compression frameworks are constructed that aim to reduce the effective number of parameters required to express the function of a trained network whilst maintaining its performance. To do this [7] capitalize on how a neural network responds to noise added to its weights. We first introduce the compression techniques for linear classifiers and then proceed to work with fully connected ReLU neural networks.

**2.3.1 ESTABLISHING THE NOTION OF COMPRESSION** We are in a scenario where we have a learned classifier  $h$  that achieves low empirical loss but is complex. In this case, we are considering  $\mathcal{Y} = \mathbb{R}^k$  so that the output of  $h$  in the  $i^{\text{th}}$  can be thought of as a relative probability that the input belongs to class  $i$ . With this, we define the classification margin loss for  $\gamma > 0$  to be

$$L_\gamma(h) = \mathbb{P}_{(x,y) \sim \mathcal{D}} \left( h(x)[y] \leq \gamma + \max_{i \neq y} f(x)[i] \right).$$

Similarly, we have the empirical classification margin loss defined as

$$\hat{L}_\gamma(h) = \frac{1}{m} \left| \left\{ f(x_i)[y_i] \leq \gamma + \max_{i \neq y_i} f(x_i)[i] \right\} \right|.$$

Suppose that our neural network has  $d$  fully connected layers and let  $x^i$  be the vector before the activation at layer  $i = 0, \dots, d$  and as  $x^0$  is the input denote it  $x$ . Let  $A^i$  be the weight matrix of layer  $i$  and let layer  $i$  have  $n_i$  hidden layers with  $n = \max_{i=1}^d n_i$ . The classifier calculated by the network will be denoted  $h_{\mathbf{w}}(x)$ , where  $\mathbf{w}$  can be thought of as a vector containing the weights of the network. For layers  $i \leq j$  the operator for the composition of the layers will be denoted  $M^{i,j}$ , the Jacobian of the input  $x$  will be denoted  $J_x^{i,j}$  and  $\phi(\cdot)$  will denote the component-wise ReLU. With these the following hold,

$$x^i = A^i \phi(x^{i-1}), \quad x^j = M^{i,j}(x^i), \quad \text{and} \quad M^{i,j}(x^i) = J_{x^i}^{i,j} x^i.$$

For a matrix  $B$ ,  $\|B\|_F$  will be its Frobenius norm,  $\|B\|_2$  its spectral norm and  $\frac{\|B\|_F^2}{\|B\|_2^2}$  its stable rank.

**Definition 2.3** Let  $h$  be a classifier and  $G_{\mathcal{W}} = \{g_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$  be a class of classifiers. We say that  $h$  is  $(\gamma, S)$ -compressible via  $G_{\mathcal{W}}$  if there exists  $\mathbf{w} \in \mathcal{W}$  such that for any  $x \in \mathcal{X}$ ,

$$|h(x)[y] - g_{\mathbf{w}}(x)[y]| \leq \gamma$$

for all  $y \in \{1, \dots, k\}$ .

**Definition 2.4** Suppose  $G_{\mathcal{W},s} = \{g_{\mathbf{w},s} : \mathbf{w} \in \mathcal{W}\}$  is a class of classifiers indexed by trainable parameters  $\mathbf{w}$  and fixed string  $s$ . A classifier  $h$  is  $(\gamma, S)$ -compressible with respect to  $G_{\mathcal{W},s}$  using helper string  $s$  if there exists  $\mathbf{w} \in \mathcal{W}$  such that for any  $x \in \mathcal{X}$ ,

$$|h(x)[y] - g_{\mathbf{w},s}(x)[y]| \leq \gamma$$

for all  $y \in \{1, \dots, k\}$ .

**Theorem 2.5** Suppose  $G_{\mathcal{W},s} = \{g_{\mathbf{w},s} : \mathbf{w} \in \mathcal{W}\}$  where  $\mathcal{W}$  is a set of  $q$  parameters each of which has at most  $r$  discrete values and  $s$  is a helper string. Let  $S$  be a training set with  $m$  samples. If the trained classifier  $h$  is  $(\gamma, S)$ -compressible via  $G_{\mathcal{W},s}$  with helper string  $s$ , then there exists  $\mathbf{w} \in \mathcal{W}$  with high probability such that

$$L_0(g_{\mathbf{w}}) \leq \hat{L}_\gamma(h) + O\left(\sqrt{\frac{q \log(r)}{m}}\right)$$

over the training set.

**Remark 2.6** Theorem 2.5 only gives a bound for  $g_{\mathbf{w}}$  which is the compression of  $h$ . However, there are no requirements of a hypothesis class, assumptions are only made on  $h$  and its properties on a finite training set.

**Corollary 2.7** If the compression works for  $1 - \xi$  fraction of the training sample, then with a high probability

$$L_0(g_{\mathbf{w}}) \leq \hat{L}_\gamma(h) + \xi + O\left(\sqrt{\frac{q \log r}{m}}\right).$$

**2.3.2 COMPRESSION OF A LINEAR CLASSIFIER** We now develop an algorithm to compress the decision vector of a linear classifier. We will use linear classifiers to conduct binary classification, where the members of one class have label 1 and the other have label  $-1$ . The linear classifiers will be parameterised by the weight vector  $\mathbf{w} \in \mathbb{R}^d$  such that for  $x \in \mathcal{X}$  we have  $h_{\mathbf{w}}(x) = \text{sgn}(\mathbf{w}^\top x)$ . Define the margin,  $\gamma > 0$ , of  $\mathbf{w}$  to be such that  $y(\mathbf{w}^\top x) \geq \gamma$  for all  $(x, y)$  in the training set. In compressing  $\mathbf{w}$ , according to Algorithm 1, we end up with a linear classifier parameterized by the weight vector  $\hat{\mathbf{w}}$  with some PAC bounds relating to its performance.

**Lemma 2.8** Algorithm 1  $(\gamma, \mathbf{w})$  returns a vector  $\hat{\mathbf{w}}$  such that for any fixed  $\mathbf{u}$ , with probability  $1 - \eta$ ,  $|\hat{\mathbf{w}}^\top \mathbf{u} - \mathbf{w}^\top \mathbf{u}| \leq \gamma$ . The vector  $\hat{\mathbf{w}}$  has at most  $O\left(\frac{\log d}{\eta \gamma^2}\right)$  non-zero entries with high probability.

In the discrete case, a similar result holds.

**Lemma 2.9** Let Algorithm 1  $(\frac{\gamma}{2}, \mathbf{w})$  return vector  $\hat{\mathbf{w}}$ . Let,

$$\hat{w}_i = \begin{cases} 0 & |\hat{w}_i| \geq 2\eta\gamma\sqrt{h} \\ \text{rounding to nearest multiple of } \frac{\gamma}{2\sqrt{h}} & \text{Otherwise.} \end{cases}$$

Then for any fixed  $\mathbf{u}$  with probability at least  $1 - \eta$ ,  $|\hat{\mathbf{w}}^\top \mathbf{u} - \mathbf{w}^\top \mathbf{u}| \leq \gamma$ .

**Algorithm 1** ( $\gamma, \mathbf{w}$ )**Require:** vector  $\mathbf{w}$  with  $\|\mathbf{w}\| \leq 1, \eta$ .**Ensure:** vector  $\hat{\mathbf{w}}$  such that for any fixed vector  $\|u\| \leq 1$ , with probability at least  $1 - \eta$ ,  $|\mathbf{w}^\top \mathbf{u} - \hat{\mathbf{w}}^\top \mathbf{u}| \leq \gamma$ .Vector  $\hat{\mathbf{w}}$  has  $O\left(\frac{\log d}{\eta \gamma^2}\right)$  non-zero entries.**for**  $i = 1 \rightarrow d$  **do**Let  $z_i = 1$  with probability  $p_i = \frac{2w_i^2}{\eta \gamma^2}$  and 0 otherwise.Let  $\hat{\mathbf{w}}_i = \frac{z_i w_i}{p_i}$ .**end for****return**  $\hat{\mathbf{w}}$ 

Using Lemmas 2.8 and 2.9 along with Corollary 2.7 the following holds.

**Lemma 2.10** *For any number of samples  $m$ , there is an efficient algorithm to generate a compressed vector  $\hat{\mathbf{w}}$ , such that*

$$L(\hat{\mathbf{w}}) \leq \tilde{O}\left(\left(\frac{1}{\gamma^2 m}\right)^{\frac{1}{3}}\right).$$

**Remark 2.11** *The rate is not optimal as it depends on  $m^{1/3}$  and not  $\sqrt{m}$ . To resolve this employ helper strings.*

**Algorithm 2** ( $\gamma, \mathbf{w}$ )**Require:** vector  $\mathbf{w}$  with  $\|\mathbf{w}\| \leq 1, \eta$ .**Ensure:** vector  $\hat{\mathbf{w}}$  such that for any fixed vector  $\|u\| \leq 1$ , with probability at least  $1 - \eta$ ,  $|\mathbf{w}^\top \mathbf{u} - \hat{\mathbf{w}}^\top \mathbf{u}| \leq \gamma$ .Let  $k = \frac{16 \log(\frac{1}{\eta})}{\gamma^2}$ .Sample the random vectors  $\mathbf{v}_1, \dots, \mathbf{v}_k \sim \mathcal{N}(0, I)$ .Let  $z_i = \langle \mathbf{v}_i, \mathbf{w} \rangle$ .(In Discrete Case) Round  $z_i$  to closes multiple of  $\frac{\gamma}{2\sqrt{dk}}$ .**return**  $\hat{\mathbf{w}} = \frac{1}{k} \sum_{i=1}^k z_i \mathbf{v}_i$ 

**Remark 2.12** *The vectors  $\mathbf{v}_i$  are the helper string.*

**Lemma 2.13** *For any fixed vector  $u$ , Algorithm 2 ( $\gamma, \mathbf{w}$ ) returns a vector  $\hat{\mathbf{w}}$  such that with probability at least  $1 - \eta$ , we have  $|\hat{\mathbf{w}}^\top \mathbf{u} - \mathbf{w}^\top \mathbf{u}| \leq \gamma$ .*

**Lemma 2.14** *For any number of sample  $m$ , there is an efficient algorithm with a helper string to generate a compressed vector  $\hat{\mathbf{w}}$ , such that*

$$L(\hat{\mathbf{w}}) \leq \tilde{O}\left(\sqrt{\frac{1}{\gamma^2 m}}\right).$$



**2.3.3 COMPRESSION OF A FULLY CONNECTED NETWORK** In a similar way, the layer matrices of a fully connected network can be compressed in such a way as to maintain a reasonable level of performance. A similar compression on how to do this is detailed in Algorithm 3. Throughout we will let  $\mathbf{w}$  parameterize our classifier. It can just be thought of as a list of layer matrices for our neural network.

---

**Algorithm 3** ( $A, \epsilon, \eta$ )

---

**Require:** Layer matrix  $A \in \mathbb{R}^{n_1 \times n_2}$ , error parameters  $\epsilon, \eta$ .

**Ensure:** Returns  $\hat{A}$  such that for all vectors  $\mathbf{u}, \mathbf{v}$  we have that

$$\mathbb{P}(|\mathbf{u}^\top \hat{A} \mathbf{v} - \mathbf{u}^\top A \mathbf{v}| \geq \epsilon \|A\|_F \|\mathbf{u}\| \|\mathbf{v}\|) \leq \eta$$

Sample  $k = \frac{\log(\frac{1}{\eta})}{\epsilon^2}$  random matrices  $M_1, \dots, M_k$  with i.i.d entries  $\pm 1$ .

**for**  $k' = 1 \rightarrow k$  **do**

    Let  $Z_{k'} = \langle A, M_{k'} \rangle M_{k'}$

**end for**

**return**  $\hat{A} = \frac{1}{k} \sum_{k'=1}^k Z_{k'}$

---

**Definition 2.15** If  $M$  is a mapping from real-valued vectors to real-valued vectors, and  $\mathcal{N}$  is a noise distribution. Then the noise sensitivity of  $M$  at  $\mathbf{x}$  with respect to  $\mathcal{N}$  is

$$\psi_{\mathcal{N}}(M, \mathbf{x}) = \mathbb{E} \left( \frac{\|M(\mathbf{x} + \eta \|\mathbf{x}\|) - M(\mathbf{x})\|^2}{\|M(\mathbf{x})\|^2} \right),$$

and let  $\psi_{\mathcal{N}, S}(M) = \max_{\mathbf{x} \in S} \psi_{\mathcal{N}}(M, \mathbf{x})$ .

**Remark 2.16** When  $\mathbf{x} \neq \mathbf{0}$  and the noise distribution is the Gaussian distribution  $\mathcal{N}(0, I)$  then the noise sensitivity of matrix  $M$  is exactly  $\frac{\|M\|_F^2}{\|M\mathbf{x}\|^2}$ . Hence, it is at most the stable rank of  $M$ .

**Definition 2.17** The layer cushion of layer  $i$  is defined as the largest  $\mu_i$  such that for any  $\mathbf{x} \in \mathcal{X}$

$$\mu_i \|A^i\|_F \|\phi(\mathbf{x}^{i-1})\| \leq \|A^i \phi(\mathbf{x}^{i-1})\|.$$

**Remark 2.18** Note that  $\frac{1}{\mu_i^2}$  is equal to the noise sensitivity of  $A^i$  at  $\phi(\mathbf{x}^{i-1})$  with respect to noise  $\eta \sim \mathcal{N}(0, I)$ .

**Definition 2.19** For layers  $i \leq j$  the inter-layer cushion  $\mu_{i,j}$  is the largest number such that

$$\mu_{i,j} \|J_{\mathbf{x}^i}^{i,j}\|_F \|\mathbf{x}^i\| \leq \|J_{\mathbf{x}^i}^{i,j} \mathbf{x}^i\|$$

for any  $\mathbf{x} \in \mathcal{X}$ . Furthermore, let  $\mu_{i \rightarrow} = \min_{i \leq j \leq d} \mu_{i,j}$ .

**Remark 2.20** Note that  $J_{x^i}^{i,i} = I$  so that

$$\frac{\|J_{x^i}^{i,i} x^i\|}{\|J_{x^i}^{i,j}\|_F \|x^i\|} = \frac{1}{\sqrt{h^i}}.$$

Furthermore,  $\frac{1}{\mu_{i,j}^2}$  is the noise sensitivity of  $J_x^{i,j}$  with respect to noise  $\eta \sim \mathcal{N}(0, I)$ .

**Definition 2.21** The activation contraction  $c$  is defined as the smallest number such that for any layer  $i$

$$\|\phi(x^i)\| \geq \frac{\|x^i\|}{c}$$

for any  $x \in \mathcal{X}$ .

**Definition 2.22** Let  $\eta$  be the noise generated as a result of applying Algorithm 3 to some of the layers before layer  $i$ . Define the inter-layer smoothness  $\rho_\delta$  to be the smallest number such that with probability  $1 - \delta$  we have that for layers  $i < j$

$$\|M^{i,j}(x^i + \eta) - J_{x^i}^{i,j}(x^i + \eta)\| \leq \frac{\|\eta\| \|x^j\|}{\rho_\delta \|x^i\|}$$

for any  $x \in \mathcal{X}$ .

**Lemma 2.23** For any  $0 < \delta$  and  $\epsilon \leq 1$  let  $G = \{(U^i, x^i)\}_{i=1}^m$  be a set of matrix-vector pairs of size  $m$  where  $U \in \mathbb{R}^{k \times n_1}$  and  $x \in \mathbb{R}^{n_2}$ , let  $\hat{A} \in \mathbb{R}^{n_1 \times n_2}$  be the output of Algorithm 3  $(A, \epsilon, \eta = \frac{\delta}{mk})$ . With probability at least  $1 - \delta$  we have for any  $(U, x) \in G$  that  $\|U(\hat{A} - A)x\| \leq \epsilon \|A\|_F \|U\|_F \|x\|$ .

**Remark 2.24** For a neural network let  $x$  be the input,  $A$  be the layer matrix and  $U$  the Jacobian of the network output with respect to the layer input. Then the network output before compression is given by  $UAx$  and after compression the output is given by  $U\hat{A}x$ .

Use Lemma 2.23 to prove the following.

**Lemma 2.25** For any fully connected network  $h_w$  with  $\rho_\delta \geq 3d$ , any probability  $0 < \delta \leq 1$  and any  $0 < \epsilon \leq 1$ , Algorithm 3 can generate weights  $\tilde{w}$  for a network with

$$\frac{72c^2 d^2 \log\left(\frac{mdn}{\delta}\right)}{\epsilon^2} \cdot \sum_{i=1}^d \frac{1}{\mu_i^2 \mu_{i \rightarrow}^2}$$

total parameters such that with probability  $1 - \frac{\delta}{2}$  over the generated weights  $\tilde{w}$ , for any  $x \in \mathcal{X}$

$$\|h_w(x) - h_{\tilde{w}}(x)\| \leq \epsilon \|h_w(x)\|,$$

where  $\mu_i, \mu_{i \rightarrow}, c$  and  $\rho_\delta$  are the layer cushion, inter-layer cushion, activation contraction and inter-layer smoother for the network.

Subsequently, the following result holds.

**Lemma 2.26** *For any fully connected network  $h_{\mathbf{w}}$  with  $\rho_\delta \geq 3d$ , any probability  $0 < \delta \leq 1$  and any margin  $\gamma > 0$ ,  $h_{\mathbf{w}}$  can be compressed (with respect to a random string) to another fully connected network  $h_{\tilde{\mathbf{w}}}$  such that for  $x \in \mathcal{X}$ ,  $\hat{L}_0(h_{\tilde{\mathbf{w}}}) \leq \hat{L}_\gamma(h_{\mathbf{w}})$  and the number of parameters in  $h_{\tilde{\mathbf{w}}}$  is at most*

$$\tilde{O}\left(\frac{c^2 d^2 \max_{x \in \mathcal{X}} \|h_{\mathbf{w}}(x)\|_2^2}{\gamma^2} \sum_{i=1}^d \frac{1}{\mu_i^2 \mu_{i \rightarrow}^2}\right).$$

**Lemma 2.27** *Let  $h_{\mathbf{w}}$  be a  $d$ -layer network with weights  $A = \{A^1, \dots, A^d\}$ . Then for any input  $x$ , weights  $A$  and  $\hat{A}$ , if for any layer  $i$ ,  $\|A^i - \hat{A}^i\| \leq \frac{1}{d} \|A^i\|$ , then*

$$\|h_{\mathbf{w}}(x) - h_{\tilde{\mathbf{w}}}(x)\| \leq e \|x\| \left( \prod_{i=1}^d \|A^i\|_2 \right) \sum_{i=1}^d \frac{\|A^i - \hat{A}^i\|_2}{\|A^i\|_2}$$

Using Lemmas 2.25, 2.26 and 2.27 it can be shown that Algorithm 3 provides a generalization bound for suitable fully connected neural networks.

**Theorem 2.28** *For any fully connected network  $h_{\mathbf{w}}$  with  $\rho_\delta \geq 3d$ , any probability  $0 < \delta \leq 1$  and any margin  $\gamma$ . Algorithm 3 generates weights  $\tilde{\mathbf{w}}$  such that with probability  $1 - \delta$  over the training set,*

$$L_0(h_{\tilde{\mathbf{w}}}) \leq \hat{L}_\gamma(h_{\mathbf{w}}) + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{x \in \mathcal{S}} \|h_{\mathbf{w}}(x)\|_2^2 \sum_{i=1}^d \frac{1}{\mu_i^2 \mu_{i \rightarrow}^2}}{\gamma^2 m}}\right).$$

### 3 EMPIRICAL PAC-BAYES BOUNDS

#### 3.1 INTRODUCTION TO PAC-BAYES THEORY

**3.1.1 BAYESIAN MACHINE LEARNING** Here we will outline an introduction to Bayesian learning given by [8]. This will provide some context to the framework under which PAC-Bayes bounds are derived. As before we suppose that our training data  $S_m = \{(x_i, y_i)\}_{i=1}^m$  consists of samples from the distribution  $\mathcal{D}$  defined on  $\mathcal{Z}$ . Bayesian machine learning is used to find a parameter  $\hat{\mathbf{w}}$  that corresponds to a hypothesis  $h_{\hat{\mathbf{w}}}$  with the property that  $h_{\hat{\mathbf{w}}}(x) \approx y$ . To do this a learning algorithm is employed, which is simply a map from the data space to the parameter space,  $\mathcal{W}$ . In this framework, our loss function takes the form  $l : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ . The learning algorithm requires some prior distribution,  $\pi$ , to be defined on  $\mathcal{W}$ . Then using the training data the posterior distribution,  $\rho$ , is formed from the prior distribution. From the posterior distribution, there are many methodologies to then determine the parameter  $\hat{\mathbf{w}}$ . For example, one could take  $\hat{\mathbf{w}}$  to be the mean, median or a random realisation of  $\rho$ .

**3.1.2 INTRODUCING PAC-BAYES BOUNDS** Bayesian machine learning is a way to manage randomness and uncertainty in the learning task. PAC-Bayes are PAC bounds that operate under this framework.

**Definition 3.1** Let  $\mathcal{M}(\mathcal{W})$  be a set of probability distributions defined over  $\mathcal{W}$ . A data-dependent probability measure is a function

$$\hat{\rho} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{M}(\mathcal{W}).$$

For ease of notation we will simple write  $\hat{\rho}$  to mean  $\hat{\rho}((X_1, Y_1), \dots, (X_n, Y_n))$ . The Kullback-Liebler (KL) divergence is a measure of similarity between probability measures defined on the same measurable space.

**Definition 3.2** Given two probability measures  $Q$  and  $P$  defined on some sample space  $\mathcal{X}$ , the KL divergence between  $Q$  and  $P$  is

$$\text{KL}(Q, P) = \int \log \left( \frac{dQ(x)}{dP(x)} \right) Q(dx)$$

when  $Q$  is absolutely continuous with respect to  $P$ . Otherwise,  $\text{KL}(Q, P) = \infty$ .

**Remark 3.3** When  $Q, P$  are probability measures on Euclidean space  $\mathbb{R}^d$  with densities  $q, p$  respectively. The KL divergence is

$$\text{KL}(Q, P) := \int \log \left( \frac{q(x)}{p(x)} \right) q(x) dx.$$

Note that KL divergence can take values in the range  $[0, \infty]$ . Also, note the asymmetry in the definition.

For the multivariate normal distributions  $N_q \sim \mathcal{N}(\mu_q, \Sigma_q)$  and  $N_p \sim \mathcal{N}(\mu_p, \Sigma_p)$  defined on  $\mathbb{R}^d$  we have that,

$$\text{KL}(N_q, N_p) = \frac{1}{2} \left( \text{tr} \left( \Sigma_p^{-1} \Sigma_q \right) - d + (\mu_p - \mu_q)^\top \Sigma_p^{-1} (\mu_p - \mu_q) + \log \left( \frac{\det \Sigma_p}{\det \Sigma_q} \right) \right).$$

Similarly, for Bernoulli distributions  $\mathcal{B}(q) \sim \text{Bern}(q)$  and  $\mathcal{B}(p) \sim \text{Bern}(p)$  it follows that

$$\text{kl}(q, p) := \text{KL}(\mathcal{B}(q), \mathcal{B}(p)) = q \log \left( \frac{q}{p} \right) + (1 - q) \log \left( \frac{1 - q}{1 - p} \right),$$

For  $p^* \in [0, 1]$  bounds of the form  $\text{KL}(q, p^*) \leq c$  for some  $q \in [0, 1]$  and  $c \geq 0$  are of interest. Hence, we introduce the notation

$$\text{kl}^{-1}(q, c) := \sup\{p \in [0, 1] : \text{kl}(q, p) \leq c\}.$$

**Proposition 3.4** *For any probability measures  $Q$  and  $P$  it follows that  $\text{KL}(Q, P) \geq 0$  with equality if and only if  $Q$  and  $P$  are the same probability distribution.*

**3.1.3 PAC-BAYES BOUNDS** The first PAC-Bayes bounds we will encounter is known as Catoni's bound. Recall, that under the Bayesian framework, we first fix a prior distribution,  $\pi \in \mathcal{M}(\mathcal{W})$ .

**Theorem 3.5 ([12])** *For all  $\lambda > 0$ , for all  $\rho \in \mathcal{M}(\mathcal{W})$ , and  $\delta \in (0, 1)$  it follows that*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (\hat{R}(\mathbf{w})) \leq \frac{\lambda C^2}{8m} + \frac{\text{KL}(\rho, \pi) + \log \left( \frac{1}{\delta} \right)}{\lambda} \right) \geq 1 - \delta.$$

*Proof.*  $\square$

Theorem 3.5 motivates the study of the data-dependent probability measure

$$\hat{\rho}_\lambda = \underset{\rho \in \mathcal{M}(\mathcal{W})}{\text{argmin}} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (\hat{R}(\mathbf{w})) + \frac{\text{KL}(\rho, \pi)}{\lambda} \right). \quad (1)$$

**Definition 3.6** *The optimization problem defined by Equation 1 has the solution  $\hat{\rho}_\lambda = \pi_{-\lambda \hat{R}}$  given by*

$$\hat{\rho}_\lambda(d\mathbf{w}) = \frac{\exp(-\lambda \hat{R}(\mathbf{w})) \pi(d\mathbf{w})}{\mathbb{E}_{\mathbf{w} \sim \pi} (\exp(-\lambda \hat{R}(\mathbf{w})))}.$$

*This distribution is known as the Gibbs posterior.*

**Corollary 3.7** *For all  $\lambda > 0$ , and  $\delta \in (0, 1)$  it follows that*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \mathbb{E}_{\mathbf{w} \sim \hat{\rho}_\lambda} (R(\mathbf{w})) \leq \inf_{\rho \in \mathcal{M}(\mathcal{W})} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (\hat{R}(\mathbf{w})) + \frac{\lambda C^2}{8m} + \frac{\text{KL}(\rho, \pi) + \log \left( \frac{1}{\delta} \right)}{\lambda} \right) \right) \geq 1 - \delta.$$

For a learning algorithm, we noted that there are different methodologies for how the learned classifier is sampled from the posterior. In the case where consider a single random realisation of the posterior distribution, we have the following result.

**Theorem 3.8** [12] For all  $\lambda > 0$ ,  $\delta \in (0, 1)$ , and data-dependent probability measure  $\tilde{\rho}$  we have that

$$\mathbb{P}_{S \sim \mathcal{D}^m} \mathbb{P}_{\tilde{\mathbf{w}} \sim \tilde{\rho}} \left( R(\tilde{\mathbf{w}}) \leq \hat{R}(\tilde{\mathbf{w}}) + \frac{\lambda C^2}{8m} + \frac{\log \left( \frac{d\rho(\tilde{\mathbf{w}})}{d\pi(\tilde{\mathbf{w}})} \right) + \log \left( \frac{1}{\delta} \right)}{\lambda} \right) \geq 1 - \delta$$

*Proof.*  $\square$  Note that Theorem 3.5 is a bound in probability. We now state an equivalent bound that holds in expectation.

**Theorem 3.9** [12] For all  $\lambda > 0$ , and data-dependent probability measure  $\tilde{\rho}$ , we have that

$$\mathbb{E}_{S \sim \mathcal{D}^m} \mathbb{E}_{\mathbf{w} \sim \tilde{\rho}} (R(\mathbf{w})) \leq \mathbb{E}_{S \sim \mathcal{D}^m} \mathbb{E}_{\mathbf{w} \sim \tilde{\rho}} \left( \hat{R}(\mathbf{w}) + \frac{\lambda C^2}{8m} + \frac{\text{KL}(\tilde{\rho}, \pi)}{\lambda} \right).$$

**Corollary 3.10** For  $\tilde{\rho} = \hat{\rho}_\lambda$ , the following holds

$$\mathbb{E}_{S \sim \mathcal{D}^m} \mathbb{E}_{\mathbf{w} \sim \hat{\rho}_\lambda} (R(\mathbf{w})) \leq \mathbb{E}_{S \sim \mathcal{D}^m} \left( \inf_{\rho \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{\mathbf{w} \sim \rho} (\hat{R}(\mathbf{w})) + \frac{\lambda C^2}{8m} + \frac{\text{KL}(\rho, \pi)}{\lambda} \right).$$

In the results that follow we will consider the 0-1 loss. This is a measurable function  $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \{0, 1\}$  defined by  $l(y, y') = \mathbf{1}(y \neq y')$ .

**Theorem 3.11** [1] For all  $\rho \in \mathcal{M}(\mathcal{W})$  and  $\delta > 0$  we have that

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (R(\mathbf{w})) \leq \mathbb{E}_{\mathbf{w} \sim \rho} (\hat{R}(\mathbf{w})) + \sqrt{\frac{\text{KL}(\rho, \pi) \log \left( \frac{1}{\delta} \right) + \frac{5}{2} \log(m) + 8}{2m - 1}} \right) \geq 1 - \delta.$$

**Theorem 3.12 ([4])** For  $a > 0$  and  $p \in (0, 1)$  let

$$\Phi_a(p) = \frac{-\log(1 - p(1 - \exp(-a)))}{a}.$$

Then for any  $\lambda > 0$ ,  $\delta > 0$  and  $\rho \in \mathcal{M}(\mathcal{W})$  we have that

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (R(\mathbf{w})) \leq \Phi_{\frac{\lambda}{m}}^{-1} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (\hat{R}(\mathbf{w})) + \frac{\text{KL}(\rho, \pi) + \log \left( \frac{1}{\delta} \right)}{\lambda} \right) \right) \geq 1 - \delta.$$

**Theorem 3.13 ([3])** For any  $\delta > 0$  and  $\rho \in \mathcal{M}(\mathcal{W})$  then we have that

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (R(\mathbf{w})) \leq \text{kl}^{-1} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (\hat{R}(\mathbf{w})), \frac{\text{KL}(\rho, \pi) + \log \left( \frac{2\sqrt{m}}{\delta} \right)}{m} \right) \right) \geq 1 - \delta.$$

### 3.2 OPTIMIZING PAC-BAYES BOUNDS VIA SGD

In practice, it is often the case that these bounds are not useful. Despite providing insight into how generalization relates to each of the components of the learning process they do not have much utility in providing non-vacuous bounds on the performance of neural networks on the underlying distribution. The significance of the KL divergence between the posterior and the prior can be noted in each of the bounds of Section 3.1.2. This motivated the work of [6] who successfully minimized this term to provide non-vacuous results in practice. They considered a restricted problem that lends itself to efficient optimization. They use stochastic gradient descent to refine the prior, which is effective as SGD is known to find flat minima. This is important as around flat minima such as  $\mathbf{w}^*$  we have that  $\hat{R}(\mathbf{w}) \approx \hat{R}(\mathbf{w}^*)$  [12]. The setup considered by [6] is the same as the one we have considered throughout this report. With  $X \subset \mathbb{R}^k$  and labels being  $\pm 1$ . That is, we are considering binary classification based on a set of features. We explicitly state our hypothesis set as

$$\mathcal{H} = \{h_{\mathbf{w}} : \mathbb{R}^k \rightarrow \mathbb{R} : \mathbf{w} \in \mathbb{R}^d\}.$$

We are still considering the 0-1, however, because our classifiers output real numbers we modify the loss slightly to account for this. That is, we let  $l : \mathbb{R} \rightarrow \{\pm 1\}$  be defined as  $l(y, y') = \mathbf{1}(\text{sgn}(y') = y)$ . For optimization purposes we use the convex surrogate loss function  $\tilde{l} : \mathbb{R} \times \{\pm 1\} \rightarrow \mathbb{R}_+$

$$\tilde{l}(y, \hat{y}) = \frac{\log(1 + \exp(-\hat{y}y))}{\log(2)}.$$

For the empirical risk under the convex surrogate loss we write

$$\tilde{R}(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m \tilde{l}(h_{\mathbf{w}}(x_i), y_i).$$

Recall, that this definition implicitly depends on the training sample  $S_m$ . We now introduce some new notations to simplify the upcoming bounds. For a randomized classifier  $\rho$  on  $\mathcal{W}$  we define the expected risk and the expected empirical risk as

$$R(\rho) = \mathbb{E}_{\mathbf{w} \sim \rho}(R(\mathbf{w})), \text{ and } \hat{R}(\rho) = \mathbb{E}_{\mathbf{w} \sim \rho}(\hat{R}(\mathbf{w}))$$

respectively. We will use these definitions throughout the remaining sections of this report as well. As noted previously the work [6] looks to minimize the KL divergence between the prior and the posterior to achieve non-vacuous bounds. To do this they work under a restricted setting and construct a process to find the posterior  $\rho$  that minimizes the divergence. To being [6] utilize the following bound.

**Theorem 3.14 ([6])** *For every  $\delta > 0, m \in \mathbb{N}$ , distribution  $\mathcal{D}$  on  $\mathbb{R}^k \times \{\pm 1\}$ , distribution  $\pi$  on  $\mathcal{W}$  and distribution  $\rho \in \mathcal{M}(\mathcal{W})$ , we have that*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \text{kl}(\hat{R}(\rho), R(\rho)) \leq \frac{\text{KL}(\rho, \pi) + \log\left(\frac{m}{\delta}\right)}{m-1} \right) \geq 1 - \delta.$$

**Remark 3.15** *Note how this is a slightly weaker statement than Theorem 3.13. This is because [6] cited this Theorem from [2], however, since then [3] was able to tighten the result by providing Theorem 3.13. In the following we will update the work of [10] and use the tightened result provided by 3.13.*

This motivates the following PAC-Bayes learning algorithm.

- Fix a  $\delta > 0$  and a distribution  $\pi$  on  $\mathcal{W}$ ,
- Collect an i.i.d sample  $S_m$  of size  $m$ ,
- Compute the optimal distribution  $\rho$  on  $\mathcal{W}$  that minimizes

$$\text{kl}^{-1} \left( \hat{R}(\rho), \frac{\text{KL}(\rho, \pi) + \log \left( \frac{2\sqrt{m}}{\delta} \right)}{m} \right), \quad (2)$$

- Then return the randomized classifier given by  $\rho$ .

Implementing such an algorithm in this general form is intractable in practice. Recall, that we are considering neural networks and so  $\mathbf{w}$  represents the weights and biased of our neural network. To make the algorithm more practical we therefore consider

$$\mathcal{M}(\mathcal{W}) = \{ \mathcal{N}_{\mathbf{w}, \mathbf{s}} = \mathcal{N}(\mathbf{w}, \text{diag}(\mathbf{s})) : \mathbf{w} \in \mathbb{R}^d, \mathbf{s} \in \mathbb{R}_+^d \}.$$

Utilizing the bound  $\text{kl}^{-1}(q, c) \leq q + \sqrt{\frac{c}{2}}$  in Equation (2) and replacing the loss with the convex surrogate loss we obtain the updated optimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^d, \mathbf{s} \in \mathbb{R}_+^d} \tilde{R}(\mathcal{N}_{\mathbf{w}, \mathbf{s}}) + \sqrt{\frac{\text{KL}(\mathcal{N}_{\mathbf{w}, \mathbf{s}}, \pi) + \log \left( \frac{2\sqrt{m}}{\delta} \right)}{2m}}. \quad (3)$$

We now suppose our prior  $\pi$  is of the form  $\mathcal{N}(\mathbf{w}_0, \lambda I)$ . As we will see the choice of  $\mathbf{w}_0$  is not too impactful, as long as it is not  $\mathbf{0}$ . However, to efficiently choose a judicious value for  $\lambda$  we discretize the problem, with the side-effect of expanding the eventual generalization bound. We let  $\lambda$  have the form  $c \exp \left( -\frac{j}{b} \right)$  for  $j \in \mathbb{N}$ , so that  $c$  is an upper bound and  $b$  controls precision. By ensuring that Theorem 3.13 holds with probability  $1 - \frac{6\delta}{\pi^2 j^2}$  for each  $j \in \mathbb{N}$  then we can apply a union bound argument to ensure that we get results that hold for probability  $1 - \delta$ . A union bound argument refers to applying Theorem 3.16.

**Theorem 3.16 ([6])** *Let  $E_1, E_2, \dots$  be events. Then  $\mathbb{P}(\bigcup_n E_n) \leq \sum_n \mathbb{P}(E_n)$ .*



Treating  $\lambda$  as continuous during the optimization process and then discretizing at the point of evaluating the bound yields the updated optimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^d, \mathbf{s} \in \mathbb{R}_+^d, \lambda \in (0, c)} \tilde{R}(\mathcal{N}_{\mathbf{w}, \mathbf{s}}) + \sqrt{\frac{1}{2} B_{\text{RE}}(\mathbf{w}, \mathbf{s}, \lambda; \delta)} \quad (4)$$

where

$$B_{\text{RE}}(\mathbf{w}, \mathbf{s}, \lambda; \delta) = \frac{\text{KL}(\mathcal{N}_{\mathbf{w}, \mathbf{s}}, \mathcal{N}(\mathbf{w}_0, \lambda I)) + 2 \log(b \log(\frac{c}{\lambda})) + \log\left(\frac{\pi^2 \sqrt{m}}{3\delta}\right)}{m}.$$

To optimize Equation (4) we would like to compute its gradient and apply SGD. However, this is not feasible in practice for  $\tilde{R}(\mathcal{N}_{\mathbf{w}, \mathbf{s}})$ . Instead we compute the gradient of  $\tilde{R}(\mathbf{w} + \xi \odot \sqrt{\mathbf{s}})$  where  $\xi \sim \mathcal{N}_{0, 1_d}$ . Once good candidates for this optimization problem are found we return to (2) to calculate the final error bound. With the choice of  $\lambda$  it follows that with probability  $1 - \delta$ , uniformly over all  $\mathbf{w} \in \mathbb{R}^d, \mathbf{s} \in \mathbb{R}_+^d$  and  $\lambda$  (of the discrete form) the expected risk of  $\rho = \mathcal{N}_{\mathbf{w}, \mathbf{s}}$  is bounded by

$$\text{kl}^{-1}(\hat{R}(\rho), B_{\text{RE}}(\mathbf{w}, \mathbf{s}, \lambda; \delta)).$$

However, it is often not possible to compute  $\hat{R}(\rho)$  due to the intractability of  $\rho$ . So instead an unbiased estimate is obtained by estimating  $\rho$  using a Monte Carlo approximation. Given  $n$  i.i.d samples  $\mathbf{w}_1, \dots, \mathbf{w}_n$  from  $\rho$  we use the Monte Carlo approximation  $\hat{\rho}_n = \sum_{i=1}^n \delta_{\mathbf{w}_i}$ , to get the bound

$$\hat{R}(\rho) \leq \overline{\hat{R}_{n, \delta'}(\rho)} := \text{kl}^{-1}\left(\hat{R}(\hat{\rho}_n), \frac{1}{n} \log\left(\frac{2}{\delta'}\right)\right),$$

which holds with probability  $1 - \delta'$ . Finally, by Theorem 3.16

$$R(\rho) \leq \text{kl}^{-1}\left(\overline{\hat{R}_{n, \delta'}(\rho)}, B_{\text{RE}}(\mathbf{w}, \mathbf{s}, \lambda; \delta)\right),$$

holds with probability  $1 - \delta - \delta'$ . Now all that is left is to do is to determine optimal values for  $\mathbf{w}$  and  $\mathbf{s}$ . To do this first train a neural network via SGD to get a value of  $\mathbf{w}$ . Then instantiate a stochastic neural network with the multivariate normal distribution  $\rho = \mathcal{N}_{\mathbf{w}, \mathbf{s}}$  over the weights, with  $\mathbf{s} = |\mathbf{w}|$ . Next apply Algorithm 4 to deduce values of  $\mathbf{w}, \mathbf{s}$  and  $\lambda$  that give a tighter bound.

Once the values of  $\mathbf{w}, \mathbf{s}$  and  $\lambda$  are found we then need to compute  $\overline{\hat{R}_{n, \delta'}(\rho)} := \text{kl}^{-1}\left(\hat{R}(\hat{\rho}_n), \frac{1}{n} \log\left(\frac{2}{\delta'}\right)\right)$  to get our bound. We note that

$$\hat{R}(\hat{\rho}_n) = \sum_{i=1}^n \delta_{\mathbf{w}_i} \left( \frac{1}{m} \sum_{j=1}^m l(h_{\mathbf{w}_i}(x_j), y_j) \right).$$

Then to invert the kl divergence we employ Newton's method, in the form of Algorithm 5, to get an approximation for our bound.

**Algorithm 4** Optimizing the PAC Bounds**Require:** $\mathbf{w}_0 \in \mathbb{R}^d$ , the network parameters at initialization. $\mathbf{w} \in \mathbb{R}^d$ , the network parameters after SGD. $S_m$ , training examples. $\delta \in (0, 1)$ , confidence parameter. $b \in \mathbb{N}, c \in (0, 1)$ , precision and bound for  $\lambda$ . $\tau \in (0, 1), T$ , learning rate.**Ensure:** Optimal  $\mathbf{w}, \mathbf{s}, \lambda$ . $\zeta = |\mathbf{w}|$  $\triangleright \mathbf{s}(\zeta) = e^{2\zeta}$  $\rho = -3$  $\triangleright \lambda(\rho) = e^{2\rho}$  $B(\mathbf{w}, \mathbf{s}, \lambda, \mathbf{w}') = \tilde{R}(\mathbf{w}) + \sqrt{\frac{1}{2} B_{\text{RE}}(\mathbf{w}, \mathbf{s}, \lambda)}$ **for**  $t = 1 \rightarrow T$  **do**Sample  $\xi \sim \mathcal{N}(0, I_d)$  $\mathbf{w}'(\mathbf{w}, \zeta) = \mathbf{w} + \xi \odot \sqrt{\mathbf{s}(\zeta)}$ 

$$\begin{pmatrix} \mathbf{w} \\ \zeta \\ \rho \end{pmatrix} = -\tau \begin{pmatrix} \nabla_{\mathbf{w}} B(\mathbf{w}, \mathbf{s}(\zeta), \lambda(\rho), \mathbf{w}'(\mathbf{w}, \zeta)) \\ \nabla_{\zeta} B(\mathbf{w}, \mathbf{s}(\zeta), \lambda(\rho), \mathbf{w}'(\mathbf{w}, \zeta)) \\ \nabla_{\rho} B(\mathbf{w}, \mathbf{s}(\zeta), \lambda(\rho), \mathbf{w}'(\mathbf{w}, \zeta)) \end{pmatrix}$$

**end for****return**  $\mathbf{w}, \mathbf{s}(\zeta), \lambda(\rho)$ **Algorithm 5** Newton's Method for Inverting kl Divergence**Require:**  $q, c$ , initial estimate  $p_0$  and  $N \in \mathbb{N}$ **Ensure:**  $p$  such that  $p \approx \text{kl}^{-1}(q, c)$ **for**  $n = 1 \rightarrow N$  **do****if**  $p \geq 1$  **then****return** 1**else**

$$p_0 = p_0 - \frac{q \log\left(\frac{q}{c}\right) + (1-q) \log\left(\frac{1-q}{1-c}\right) - c}{\frac{1-q}{1-p} - \frac{q}{p}}$$

**end if****end for****return**  $p_0$

## 4 ORACLE PAC-BAYES BOUNDS

### 4.1 THEORY OF ORACLE PAC-BAYES BOUNDS

Oracle bounds are theoretical objects that are not suitable for practical applications. Their utility lies in their ability to highlight properties about the behaviour of our bounds. For example, they can take the form

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( R(\hat{\mathbf{w}}) \leq \inf_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w}) + r_m(\delta) \right) \geq 1 - \delta.$$

Where  $r_m(\delta)$  is a remainder term that tends to 0 as  $m$  tends to  $\infty$ . Although this bound cannot be computed in practice it is illustrative of the behaviour of the bound. Just like empirical bounds, there exist oracle bounds that hold in expectation and in probability.

#### 4.1.1 ORACLE PAC-BAYES BOUNDS IN EXPECTATION

**Theorem 4.1** For  $\lambda > 0$  we have that

$$\mathbb{E}_{S \sim \mathcal{D}^m} \mathbb{E}_{\mathbf{w} \sim \hat{\rho}_\lambda} (R(\mathbf{w})) \leq \inf_{\rho \in \mathcal{M}(\mathcal{W})} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (R(\mathbf{w})) + \frac{\lambda C^2}{8m} + \frac{\text{KL}(\rho, \pi)}{\lambda} \right).$$

*Proof.*  $\square$

#### 4.1.2 ORACLE PAC-BAYES BOUNDS IN PROBABILITY

**Theorem 4.2** For any  $\lambda > 0$ , and  $\delta \in (0, 1)$  we have that

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( \mathbb{E}_{\mathbf{w} \sim \hat{\rho}_\lambda} (R(\mathbf{w})) \leq \inf_{\rho \in \mathcal{M}(\mathcal{W})} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (R(\mathbf{w})) + \frac{\lambda C^2}{4m} + \frac{2\text{KL}(\rho, \pi) + \log(\frac{2}{\delta})}{\lambda} \right) \right) \geq 1 - \delta$$

*Proof.*  $\square$

#### 4.1.3 BERNSTEIN'S ASSUMPTION

**Definition 4.3** Let  $\mathbf{w}^*$  denote a minimizer of  $R$  when it exists,

$$R(\mathbf{w}^*) = \min_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w}).$$

When  $\mathbf{w}^*$  exists and there is a constant  $K$  such that for any  $\mathbf{w} \in \mathcal{W}$  we have that

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left( (l(\mathbf{w}(X_i), y_i) - l(\mathbf{w}^*(X_i), y_i))^2 \right) \leq K (R(\mathbf{w}) - R(\mathbf{w}^*))$$

we say that Bernstein's assumption is satisfied with constant  $K$ .

**Theorem 4.4** Assume Bernstein's assumption is satisfied with some constant  $K > 0$ . Take  $\lambda = \frac{m}{\max(2K, C)}$  then we have

$$\mathbb{E}_{S \sim \mathcal{D}^m} \mathbb{E}_{\mathbf{w} \sim \hat{\rho}_\lambda} (R(\mathbf{w})) - R(\mathbf{w}^*) \leq 2 \inf_{\rho \in \mathcal{M}(\mathcal{W})} \left( \mathbb{E}_{\mathbf{w} \sim \rho} (R(\mathbf{w})) - R(\mathbf{w}^*) + \frac{\max(2K, C) \text{KL}(\rho, \pi)}{m} \right).$$

*Proof.*  $\square$

## 4.2 DATA DRIVEN PAC-BAYES BOUNDS

A lot of work to obtain non-vacuous PAC-Bayes bounds is to develop priors that reduce the size of the KL divergence between the prior and the posterior. The idea behind the work of [10] is to hold out some of the training data to obtain data-inspired priors. For this section, we use a PAC-Bayes bound that can be thought of as the Bayesian equivalent of Theorem 2.2, however, now we are dealing with potentially uncountable hypothesis sets.

**Theorem 4.5 ([5])** *For  $\lambda > \frac{1}{2}$  selected before drawing our training sample, then for all  $\rho \in \mathcal{M}(\mathcal{W})$  and  $\delta \in (0, 1)$  we have that*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( R(\rho) \leq \hat{R}(\rho) + \frac{\lambda C}{m} \left( \text{KL}(\rho, \pi) + \log \left( \frac{1}{\delta} \right) \right) \right) \geq 1 - \delta.$$

*Proof.*  $\square$

**Corollary 4.6 ([10])** *Let  $\beta, \delta \in (0, 1)$ ,  $\mathcal{D}$  a probability distribution over  $\mathcal{Z}$ , and  $\pi \in \mathcal{M}(\mathcal{W})$ . Then for all  $\rho \in \mathcal{M}(\mathcal{W})$  we have that*

$$\mathbb{P}_{S \sim \mathcal{D}^m} (R(\rho) \leq \Psi_{\beta, \delta}(\rho, \pi; S)) \geq 1 - \delta,$$

where  $\Psi_{\beta, \delta}(\rho, \pi; S) = \frac{1}{\beta} \hat{R}(\rho) + \frac{\text{KL}(\rho, \pi) + \log(\frac{1}{\delta})}{2\beta(1-\beta)m}$ .

*Proof.*  $\square$  As we have done previously, we can consider the optimization problem of minimizing the bound of Corollary 4.6.

**Theorem 4.7 ([10])** *Let  $m \in \mathbb{N}$  and fix a probability kernel  $\rho : \mathcal{Z}^m \rightarrow \mathcal{M}(\mathcal{W})$ . Then for all  $\beta, \delta \in (0, 1)$  and distributions  $\mathcal{D}$  defined on  $\mathcal{Z}$  we that  $\mathbb{E}_{S \sim \mathcal{D}^m} (\Psi_{\beta, \delta}(\rho(S), \pi; S))$  is minimized, in  $\pi$ , by the oracle prior  $\pi^* = \mathbb{E}_{S \sim \mathcal{D}^m} (\rho(S))$ .*

For a subset  $J$  of  $\{1, \dots, m\}$  of size  $n$ , we can use it to sample the training data and yield the subset  $S_J$ . We can then define the data-dependent oracle prior as

$$\pi^*(S_J) = \inf_{\pi \in \mathcal{Z}^n \rightarrow \mathcal{M}(\mathcal{W})} \mathbb{E}(\text{KL}(\rho(s), \pi(S_J)))$$

which turns out to be  $\pi^*(S_J) = \mathbb{E}(\rho(S)|S_J)$ . It can be shown that the data-dependent oracle prior minimizes the bound of Corollary 4.6 in expectation. Therefore, despite being a theoretical quantity, as it cannot be computed in practice, it motivates the construction of practical data-dependent priors as a method to tighten the bounds.

**4.2.1 IMPLEMENTING DATA-DEPENDENT PRIORS** To implement data-dependent priors we restrict the optimization problem to make it tractable. We only consider the set of Gaussian priors  $\mathcal{F}$  that generate Gaussian posteriors. Neural networks are trained via SGD, and hence there is some randomness to the learning algorithm. Let  $(\Omega, \mathcal{F}, \nu)$  define a probability space and let us focus on the kernels

$$\rho : \Omega \times \mathcal{Z}^m \rightarrow \mathcal{M}(\mathcal{W}), \quad \rho(U, S) = \mathcal{N}(\mathbf{w}_s, \mathbf{s}),$$

where  $\mathbf{w}_S$  are the learned weights via SGD on the full dataset  $S$ . The random variable  $U$  represents the randomness of the learning algorithm. As before we consider a non-negative integer  $n \leq m$  and with  $\alpha = \frac{n}{m}$  we define a subset  $S_\alpha$  of size  $n$  containing the first  $n$  indices of  $S$  processed by SGD. Let  $\mathbb{E}^{S_\alpha, U}[\cdot]$  denote the conditional expectation operator given  $S_\alpha$  and  $U$ . Our aim now is to tighten the bound of Corollary 4.6 by minimizing  $\mathbb{E}^{S_\alpha, U}(\text{KL}(\rho(U, S), \pi))$ . To do this we further restrict the priors of consideration to those of the form  $\mathcal{N}(\mathbf{w}_\alpha, \sigma I)$  such that with  $\sigma$  fixed we are left with the minimization problem

$$\text{argmin}_{\mathbf{w}_\alpha} (\mathbb{E}^{S_\alpha, U} (\|\mathbf{w}_S - \mathbf{w}_\alpha\|)), \quad (5)$$

which can be solved to yield  $\mathbf{w}_\alpha = \mathbb{E}^{S_\alpha, U}(\mathbf{w}_S)$ . This minimizer is unknown in practice so we attempt to approximate it. We first define a so-called ghost sample,  $S^G$ , which is an independent sample equal in distribution to  $S$ . We combine a  $1 - \alpha$  fraction of  $S^G$  with  $S_\alpha$  to obtain the sample  $S_\alpha^G$ . Let  $\mathbf{w}_\alpha^G$  be the mean of  $\rho(U, S_\alpha^G)$ . By construction, SGD will first process  $S_\alpha$  then the combined portion of  $S^G$  and hence  $\mathbf{w}_\alpha^G$  and  $\mathbf{w}_S$  are equal in distribution when conditioned on  $S_\alpha$  and  $U$ . Therefore,  $\mathbf{w}_\alpha^G$  is an unbiased estimator of  $\mathbb{E}^{S_\alpha, U}(\mathbf{w}_S)$ . Before formalizing this process algorithmically we clarify some notation.

- The SGD run on  $S$  is the base run.
- The SGD run on  $S_\alpha$  is the  $\alpha$ -prefix run.
- The SGD run on  $S_\alpha^G$  is the  $\alpha$ -prefix+ghost run and obtains the parameters  $\mathbf{w}_\alpha^G$ .

The resulting parameters of the  $\alpha$ -prefix and  $\alpha$ -prefix+ghost run can be used as the centres of the Gaussian priors to give the tightened generalization bounds. However, sometimes the ghost sample is not attainable in practice, and hence one simply relies upon  $\alpha$ -prefix runs to obtain the mean of the prior. It is not clear whether  $\alpha$ -prefix+ghost run will always obtain a parameter that leads to a tighter generalization bound. Recall, that  $\sigma$  is assumed to be fixed in the optimization process. Algorithm 7 is independent of this parameter and so it can be optimized afterwards without requiring a re-run of the optimization process.

**Algorithm 6** Stochastic Gradient Descent**Require:** Learning rate  $\eta$ 


---

```

function SGD( $\mathbf{w}_0, S, b, t, \mathcal{E} = -\infty$ )
   $\mathbf{w} \leftarrow \mathbf{w}_0$ 
  for  $i \leftarrow 1$  to  $t$  do
    Sample  $S' \in S$  with  $|S'| = b$ 
     $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla l_{S'}(\mathbf{w})$ 
    if  $l_S^{0-1}(\mathbf{w}) \leq \mathcal{E}$  then
      break
    end if
  end for
end function

```

---

**Algorithm 7** Obtaining Bound Using SGD Informed Prior**Require:** Stopping criteria  $\mathcal{E}$ , Prefix fraction  $\alpha$ , Ghost Data  $S^G$  (If available), Batch size  $b$ .

---

```

function GETBOUND( $\mathcal{E}, \alpha, T, \sigma_p$ )
   $S_\alpha \leftarrow \{z_1, \dots, z_{\alpha|S|} \subset S\}$ 
   $\mathbf{w}_\alpha^0 \leftarrow \text{SGD}(\mathbf{w}_0, S_\alpha, b, \frac{|S_\alpha|}{b})$ 
   $\mathbf{w}_S \leftarrow \text{SGD}(\mathbf{w}_\alpha^0, S, b, \infty, \mathcal{E})$ 
   $\mathbf{w}_\alpha^G \leftarrow \text{SGD}(\mathbf{w}_\alpha^0, S_\alpha^G, b, T, \cdot)$ 
   $\pi \leftarrow \mathcal{N}(\mathbf{w}_\alpha^G, \sigma I)$ 
   $\rho \leftarrow \mathcal{N}(\mathbf{w}_S, \sigma I)$ 
  Bound  $\leftarrow \Psi_\delta^*(\rho, \pi; S \setminus S_\alpha)$ 
  return Bound
end function

```

---

$\triangleright$  Base Run  
 $\triangleright$  Ghost run if data available, otherwise prefix run

## 5 EXTENSIONS OF PAC-BAYES BOUNDS

### 5.1 DISINTEGRATED PAC-BAYES BOUNDS

The majority of the PAC-Bayes bounds we have discussed so far have been derived to hold for all posterior distributions. The intention of disintegrated PAC-Bayes bounds is to refine these results by only requiring them to hold for a single posterior distribution. We now study the work of [11] that sets out a general framework for deriving such bounds. The setup is the same as the one we have considered so far, with the added assumption that  $C = 1$  and the additional consideration of a deterministic learning algorithm  $A : \mathcal{Z}^m \rightarrow \mathcal{M}(\mathcal{W})$  that is applied to the training sample  $S$ .

**Definition 5.1 ([11])** *The two distributions  $P$  and  $Q$  defined on the some sample space  $\mathcal{X}$ , then for any  $\alpha > 1$  their Renyi divergence is defined to be*

$$D_\alpha(Q, P) = \frac{1}{\alpha - 1} \log \left( \mathbb{E}_{x \sim P} \left( \frac{Q(x)}{P(x)} \right)^\alpha \right).$$

**Theorem 5.2 ([11])** *For any distribution  $\mathcal{D}$  on  $\mathcal{Z}$ , for any parameter space  $\mathcal{W}$ , for any prior distribution  $\pi$  on  $\mathcal{W}$ , for any  $\phi : \mathcal{W} \times \mathcal{Z}^m \rightarrow \mathbb{R}^+$ , for any  $\alpha > 1$ , for any  $\delta > 0$  and for any deterministic learning algorithm  $A : \mathcal{Z}^m \rightarrow \mathcal{M}(\mathcal{W})$  the following holds*

$$\mathbb{P}_{S \sim \mathcal{D}^m, \mathbf{w} \sim \rho_S} \left( \frac{\alpha}{\alpha - 1} \log (\phi(\mathbf{w}, S)) \leq \frac{2\alpha - 1}{\alpha - 1} \log \left( \frac{2}{\delta} \right) + D_\alpha(\rho_S, \pi) + \log \left( \mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{E}_{\mathbf{w}' \sim \pi} \phi(\mathbf{w}', S')^{\frac{\alpha}{\alpha-1}} \right) \right) \geq 1 - \delta,$$

where  $\rho_S := A(S)$ .

*Proof.*  $\square$

**5.1.1 APPLICATION TO NEURAL NETWORK CLASSIFIERS** We can contextualize this bound to over-parameterized neural networks. Suppose that  $\mathbf{w} \in \mathbb{R}^d$  is a weight vector of a neural network, with  $d \gg m$ . Assume that the network is trained for  $T$  epochs and that these epochs are used to generate  $T$  priors  $\mathbf{P} = \{\pi_t\}_{t=1}^T$ . Let the priors be of the form  $\pi_t = \mathcal{N}(\mathbf{w}_t, \sigma^2 \mathbf{I}_d)$  where  $\mathbf{w}_t$  is the weight vector obtained after the  $t^{\text{th}}$  epoch. We assume that the priors are obtained from the learning algorithm being applied to the sample  $S_{\text{prior}}$  where  $S_{\text{prior}} \cap S = \emptyset$ .

**Corollary 5.3** *For any distribution  $\mathcal{D}$  on  $\mathcal{Z}$ , for any set  $\mathcal{W}$ , for any set  $\mathbf{P}$  of  $T$  priors on  $\mathcal{W}$ , for any learning algorithm  $A : \mathcal{Z}^m \rightarrow \mathcal{M}(\mathcal{W})$ , for any loss  $l : \mathcal{W} \times \mathcal{Z} \rightarrow [0, 1]$  and for any  $\delta > 0$  then for any  $\pi_t \in \mathbf{P}$  we have that*

$$\mathbb{P}_{S \sim \mathcal{D}^m, \mathbf{w} \sim \rho_S} \left( \text{kl}(\hat{R}(\mathbf{w}), R(\mathbf{w})) \leq \frac{1}{m} \left( \frac{\|\mathbf{w} - \mathbf{w}_t\|_2^2}{\sigma^2} + \log \left( \frac{16T\sqrt{m}}{\delta^3} \right) \right) \right) \geq 1 - \delta.$$

**Corollary 5.4** *Under the assumptions of Corollary 5.3 with  $\delta \in (0, 1)$  and for all  $\pi_t \in \mathbf{P}$  we have that*

$$\mathbb{P}_{S \sim \mathcal{D}^m, \mathbf{w} \sim \rho_S} \left( \text{kl}(\hat{R}(\mathbf{w}), R(\mathbf{w})) \leq \frac{1}{m} \left( \frac{\|\mathbf{w} + \epsilon - \mathbf{w}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \log \left( \frac{2T\sqrt{m}}{\delta} \right) \right) \right),$$

$$\mathbb{P}_{S \sim \mathcal{D}^m, \mathbf{w} \sim \rho_S} \left( \text{kl}(\hat{R}(\mathbf{w}), R(\mathbf{w})) \leq \frac{1}{m} \left( \frac{m+1}{m} \frac{\|\mathbf{w} + \epsilon - \mathbf{w}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \log \left( \frac{T(m+1)}{\delta} \right) \right) \right),$$

and for all  $c \in \mathbf{C}$

$$R(\mathbf{w}) \leq \frac{1 - \exp \left( -c\hat{R}(\mathbf{w}) - \frac{1}{m} \left( \frac{\|\mathbf{w} + \epsilon - \mathbf{w}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \log \left( \frac{T|\mathbf{C}|}{\delta} \right) \right) \right)}{1 - \exp(-c)}.$$

Where  $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$  is Gaussian noise such that  $\mathbf{w} + \epsilon$  acts as the weights sampled from  $\mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$ , and  $\mathbf{C}$  is a set of hyper-parameters fixed a priori.

## 5.2 PAC-BAYES COMPRESSION BOUNDS

We will now see how compression ideas can be capitalized to tighten PAC-Bayes bounds. The work of [9] evaluates generalization bounds by first measuring the effective compressed size of a neural network and then substituting this into the bounds. We have seen that compression techniques can efficiently reduce the effective size of a network, and so accounting for this can lead to tighter bounds. This also captures the intuition that we expect a model to overfit if it is more difficult to compress. Therefore, these updated bounds also incorporate a notion of model complexity. The work of [9] utilizes a refined version of Theorem 3.12.

**Theorem 5.5 ([4])** *Let  $L$  be a 0-1 valued loss function. Let  $\pi$  be a probability measure on the parameter space, and let  $\alpha > 1, \delta > 0$ . Then,*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left( R(\rho) \leq \inf_{\lambda > 1} \Phi_{\lambda/m}^{-1} \left( \hat{R}(\rho) + \frac{\alpha}{\lambda} \left( \text{KL}(\rho, \pi) - \log(\delta) + 2 \log \left( \frac{\log(\alpha^2 \lambda)}{\log(\alpha)} \right) \right) \right) \right) \geq 1 - \delta.$$

*Proof.*  $\square$  The intention now is to motivate the choice of  $\pi$  using ideas of compressibility such that  $\text{KL}(\rho, \pi)$  is kept small. To do this we will choose a prior  $\pi$  that assigns greater probability mass to models with a shorter code length.

**Theorem 5.6 ([9])** *Let  $|h|_c$  denote the number of bits required to represent hypothesis  $h$  using some pre-specified coding  $c$ . Let  $\rho$  denote the point mass distribution at the compressed model  $\hat{h}$ . Let  $M$  denote any probability measure on the positive integers. Then there exists a prior  $\pi_c$  such that*

$$\text{KL}(\rho, \pi_c) \leq |h|_c \log(2) - \log \left( M \left( |h|_c \right) \right).$$

**Remark 5.7** *An example of a coding scheme  $c$  could be the Huffman encoding. However, such a compression scheme is agnostic to any structure of the hypothesis class  $\mathcal{H}$ . By exploiting structure in the hypothesis class the bound can be improved substantially.*



We now formalise compression schemes to allow us to refine Theorem 5.6. Denote a compression procedure by a triple  $(S, C, Q)$  where

- $S = \{s_1, \dots, s_k\} \subseteq \{1, \dots, d\}$  is the location of the non-zero weights,
- $C = \{c_1, \dots, c_r\} \subseteq \mathbb{R}$ , is a codebook, and
- $Q = (q_1, \dots, q_k)$  for  $q_i \in \{1, \dots, r\}$  are the quantized values.

Define the corresponding weights  $\mathbf{w}(S, Q, C) \in \mathbb{R}^d$  as,

$$w_i(S, Q, C) = \begin{cases} c_{q_j} & i = s_j \\ 0 & \text{otherwise.} \end{cases}$$

Training a neural network is a stochastic process due to the randomness of SGD. So to analyse the generalization error we try to capture randomness in the analysis by applying Gaussian noise to weights. For this we use  $\rho \sim \mathcal{N}(\mathbf{w}, \sigma^2 J)$ , with  $J$  being a diagonal matrix.

**Theorem 5.8 ([9])** *Let  $(S, C, Q)$  be the output of a compression scheme, and let  $\rho_{S,C,Q}$  be the stochastic estimator given by the weights decoded from the triplet and variance  $\sigma^2$ . Let  $c$  denote an arbitrary fixed coding scheme and let  $M$  denote an arbitrary distribution on the positive integers. Then for any  $\tau > 0$ , there is a prior  $\pi$  such that*

$$\begin{aligned} \text{KL}(\rho_{S,C,Q}, \pi) &\leq (k \lceil \log(r) \rceil + |S|_c + |C|_c) \log(2) - \log(M(k \lceil \log(r) \rceil + |S|_c + |C|_c)) \\ &\quad + \sum_{i=1}^k \text{KL} \left( \mathcal{N}(c_{q_i}, \sigma^2), \sum_{j=1}^r \mathcal{N}(c_j, \tau^2) \right). \end{aligned}$$

Choosing the prior alluded to by Theorem 5.8 and utilizing Theorem 5.5 one can obtain a PAC-Bayes generalization bound that exploits notions of compressibility.

### 5.3 MUTUAL INFORMATION

**REFERENCES**

- [1] David A. McAllester. “Some PAC-Bayesian Theorems”. In: *Machine Learning* 37 (1998), pp. 355–363.
- [2] John Langford and Matthias Seeger. “Bounds for Averaging Classifiers”. In: (Feb. 2001).
- [3] Andreas Maurer. “A Note on the PAC Bayesian Theorem”. In: *CoRR* (2004).
- [4] Olivier Catoni. “Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning”. In: *IMS Lecture Notes Monograph Series* 56 (2007), pp. 1–163.
- [5] David A. McAllester. “A PAC-Bayesian Tutorial with A Dropout Bound”. In: *CoRR* (2013).
- [6] Gintare Karolina Dziugaite and Daniel M. Roy. “Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data”. In: *CoRR* (2017).
- [7] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. “Stronger generalization bounds for deep nets via a compression approach”. In: *CoRR* (2018).
- [8] Benjamin Guedj. *A Primer on PAC-Bayesian Learning*. 2019.
- [9] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. *Non-Vacuous Generalization Bounds at the ImageNet Scale: A PAC-Bayesian Compression Approach*. 2019.
- [10] Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, and Daniel M. Roy. “On the role of data in PAC-Bayes bounds”. In: *CoRR* (2020).
- [11] Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. *A General Framework for the Disintegration of PAC-Bayesian Bounds*. 2021.
- [12] Pierre Alquier. *User-friendly introduction to PAC-Bayes bounds*. 2023.