A Guide to Probably Approximately Correct Bounds for Neural Networks

Thomas Walker

Supervised by Professor Alessio Lomuscio

Summer 2023

Contents

1 Introduction

A great resource for introducing the field of Probably Approximately Correct (PAC) learning theory is given in [27]. It details the progression of results in the field and motivates the various research avenues. PAC learning theory is a general framework for studying learning algorithms, and my aim here is to illustrate how this theory is being contextualized within machine learning, with a specific focus on neural networks. With this report, I want to introduce the theory and detail some applications, as well as provide some recent extensions. The main product of PAC learning theory is bounds on the performance of the output of learning algorithms, termed PAC bounds. This report will not provide an exhaustive list of the various PAC bounds being applied to neural networks. I will instead provide some well-known results in the literature and how some of them manifest in applications. For a comprehensive introduction to the field of PAC, the reader is recommended to refer to [27]. Nevertheless, this report will be mostly self-contained, with proofs for the major results and elaboration on the 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 [27] 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 concern ourselves most with will be supervised classification tasks. 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 $h : \mathcal{X} \to \mathcal{Y}$ that correctly labels samples from $\mathcal{X}$ according to $\mathcal{D}$. The training data $S = \{(x_i, y_i)\}_{i=1}^m$ consists of m i.i.d samples from $\mathcal{D}$. As we are considering neural networks, a classifier will be parameterized 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. We will often denote the set of probability distributions over $\mathcal{W}$ as $\mathcal{M}(\mathcal{W})$. To assess the quality of a classifier we define a measurable function $l : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ 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

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

As our classifier is parameterized $\mathbf{w}$ we will instead 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 $\mathbb{E}_{S\sim\mathcal{D}^m} \left(\hat{R}(\mathbf{w}) \right) = R(\mathbf{w})$.

2.1.2 PAC Bounds

PAC 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 ([27])**.** *Let $|\mathcal{W}| = M < \infty$, $\delta \in (0, 1)$, and $\mathbf{w} \in \mathcal{W}$ then it follows that*

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

Theorem 2.1.1 (Markov's Inequality)**.** *For X a non-negative random variable and $\alpha > 0$ we have that*

$$\mathbb{P}\left(X \geq \alpha\right) \leq \frac{\mathbb{E}(X)}{\alpha}.$$

Proof. Define Y as the indicator random variable $\mathbb{I}_{\{X\geq\alpha\}}$ so that $\mathbb{E}(Y) = \mathbb{P}\left(X \geq \alpha\right)$. It is clear that $\alpha Y \leq X$ which means that $\mathbb{E}(\alpha Y) \leq \mathbb{E}(X)$, which implies that $\alpha\mathbb{P}(X \geq \alpha) \leq \mathbb{E}(X)$. Using the fact that $\alpha > 0$ we can re-arrange this expressions to complete the proof of the theorem. ■

Corollary 2.1.2 (Chernoff Bound)**.** *For a random variable X, $t > 0$ and $a \in \mathbb{R}$ we have that*

$$\mathbb{P}\left(X \geq a\right) = \mathbb{E}\left(\exp(tX)\right) \exp(-ta)$$

for $t > 0$.

Proof. This follows from Markov's inequality due to the increasing, positivity and injectivity of $\exp(\cdot)$ in particular we have that

$$\mathbb{P}\left(X \geq a\right) = \mathbb{P}\left(\exp(tX) \geq e^{ta}\right) \leq \frac{\mathbb{E}\left(\exp(tX)\right)}{e^{ta}},$$

which completes the proof. ■

Lemma 2.1.3 ([11]). *Let $U_1, \ldots, U_n$ be independent random variables taking values in an interval $[a, b]$. Then for any $t > 0$ we have that*

$$\mathbb{E}\left(\exp\left(t\sum_{i=1}^{n}\left(U_i - \mathbb{E}(U_i)\right)\right)\right) \leq \exp\left(\frac{nt^2(b-a)^2}{8}\right).$$

Proof. For $s > 0$ the function $x \mapsto e^{sx}$ is convex so that

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}.$$

Let $V_i = U_i - \mathbb{E}(U_i)$, then as $\mathbb{E}(V_i) = 0$ it follows that

$$\mathbb{E}\left(\exp(sV_i)\right) \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}.$$

With $p = \frac{b}{b-a}$ and $u = (b-a)s$ consider

$$\psi(u) = \log\left(pe^{sa} + (1-p)e^{sb}\right) = (p-1)u + \log\left(p + (1-p)e^{u}\right).$$

This is a smooth function so that by Taylor's theorem we have that for any $u \in \mathbb{R}$ there exists $\xi = \xi(u) \in \mathbb{R}$ such that

$$\psi(u) = \psi(0) + \psi'(0)u + \frac{1}{2}\psi''(\xi)u^2.$$

As

$$\psi'(u) = (p-1) + 1 - \frac{p}{p + (1-p)e^{u}}$$

we have that $\psi(0) = 0$ and $\psi'(0) = 0$. Furthermore, as

$$\psi''(u) = \frac{p(1-p)e^{u}}{(p + (1-p)e^{u})^2}, \text{ and } \psi^{(3)}(u) = \frac{p(1-p)e^{u}(p + (1-p)e^{u})(p - (1-p)e^{u})}{(p + (1-p)e^{u})^2}$$

we see that $\psi''(u)$ has a stationary point at $u^* = \log\left(\frac{p}{p-1}\right)$. For u slightly less than u^* we have $\psi^{(3)}(u) > 0$ and for u slightly larger than u^* we have $\psi^{(3)}(u) < 0$. Therefore, u^* is a maximum point and so

$$\psi''(u) \leq \psi''(u^*) = \frac{1}{4}.$$

Hence, $\psi(u) \leq \frac{u^2}{8}$ which implies that

$$\log\left(\mathbb{E}\left(\exp(sV_i)\right)\right) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left(\exp\left(t\sum_{i=1}^{n}\left(U_i-\mathbb{E}(U_i)\right)\right)\right) &= \prod_{i=1}^{n}\mathbb{E}\left(\exp\left(t(U_i-\mathbb{E}(U_i))\right)\right)\\ &\leq \prod_{i=1}^{n}\exp\left(\frac{t^2(b-a)^2}{8}\right)\\ &\leq \exp\left(\frac{nt^2(b-a)^2}{8}\right)\end{aligned}$$

which completes the proof. ■

Proof. Recall that we have our random sample $S = \{(x_i, y_i)\}_{i=1}^{m} \sim \mathcal{D}^m$. If we fix $\mathbf{w} \in \mathcal{W}$ we can let $l_i(\mathbf{w}) = l(h_{\mathbf{w}}(x_i), y_i)$. This is a random variable due to the randomness of S and so we can apply Lemma 2.1.3 to $U_i = \mathbb{E}(l_i(\mathbf{w})) - l_i(\mathbf{w})$ to get that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(tm\left(R(\mathbf{w})-\hat{R}(\mathbf{w})\right)\right)\right) \leq \exp\left(\frac{mt^2C^2}{8}\right).$$

Therefore, for any $s > 0$ we can apply Markov's Inequality to get that

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\mathbf{w})-\hat{R}(\mathbf{w}) > s\right) &= \mathbb{P}_{S\sim\mathcal{D}^m}\left(\exp\left(mt\left(R(\mathbf{w})-\hat{R}(\mathbf{w})\right)\right) > \exp(mts)\right)\\ &\leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(mt\left(R(\mathbf{w})-\hat{R}(\mathbf{w})\right)\right)\right)}{\exp(mts)}\\ &\leq \exp\left(\frac{mt^2C^2}{8}-mts\right).\end{aligned}$$

This bound is minimized for $t = \frac{4s}{C^2}$ so that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\mathbf{w}) > \hat{R}(\mathbf{w}) + s\right) \leq \exp\left(-\frac{2ms^2}{C^2}\right).$$

The above bound holds for fixed $\mathbf{w} \in \mathcal{W}$ so develop a uniform bound we consider the following.

$$\begin{aligned}\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left(\sup_{\mathbf{w}\in\mathcal{W}}\left(R(\mathbf{w})-\hat{R}(\mathbf{w})\right) > s\right) &= \mathbb{P}_{S\sim\mathcal{D}^m}\left(\bigcup_{\mathbf{w}\in\mathcal{W}}\left\{R(\mathbf{w})-\hat{R}(\mathbf{w}) > s\right\}\right)\\ &\leq \sum_{\mathbf{w}\in\mathcal{W}}\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\mathbf{w}) > \hat{R}(\mathbf{w}) + s\right)\\ &\leq M\exp\left(-\frac{2ms^2}{C^2}\right).\end{aligned}$$

Now taking $\delta = M\exp\left(-\frac{2ms^2}{C^2}\right)$ we get that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\sup_{\mathbf{w}\in\mathcal{W}}\left(R(\mathbf{w})-\hat{R}(\mathbf{w})\right) > C\sqrt{\frac{\log\left(\frac{M}{\delta}\right)}{2m}}\right) \leq \delta$$

which upon taking complements completes the proof of the theorem. □

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 ϵ 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 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 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 [24]. Uniform convergence bounds have the general form

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

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

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

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 π 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 π is defined over $\mathcal{W}$.

Theorem 2.2 ([10]). *Simultaneously for all $\mathbf{w}\in\mathcal{W}$ and $\delta\in(0,1)$ the following holds,*

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

Theorem 2.2.1 (Relative Chernoff Bound 1 [5]). *Suppose $X_1,\ldots,X_n$ are independent random variables with range $\{0,1\}$. Let $\mu=\sum_{i=1}^n X_i$. Then for $\delta\in(0,1)$ we have*

$$\mathbb{P}\left(X\leq(1-\delta)\mu\right)\leq\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}.$$

Proof. Using Markov's inequality we note that for $t<0$ we have

$$\begin{aligned}\mathbb{P}\left(X\leq(1-\delta)\mu\right)&=\mathbb{P}\left(e^{tX}\geq e^{t(1-\delta)\mu}\right)\\&\leq\frac{\mathbb{E}\left(e^{tX}\right)}{e^{t(1-\delta)\mu}}\\&\leq\frac{\exp\left(\left(e^t-1\right)\mu\right)}{\exp\left(t(1-\delta)\mu\right)}.\end{aligned}$$

Setting $t=\log(1-\delta)<0$ we get that

$$\mathbb{P}\left(X\leq(1-\delta)\mu\right)\leq\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.$$

which completes the proof of the theorem. ■

Corollary 2.2.2 ([5]). *Suppose $X_1, \ldots, X_n$ are independent random variables with range $\{0,1\}$. Let $\mu = \sum_{i=1}^n X_i$. Then for $\delta \in (0,1)$ we have*

$$\mathbb{P}\left(X \leq (1-\delta)\mu\right) \leq \exp\left(-\frac{\mu\delta^2}{2}\right).$$

Proof. Consider

$$f(\delta) = -\delta - (1-\delta)\log(1-\delta) + \frac{\delta^2}{2}$$

for $\delta \in (0,1)$. Note that

$$f'(\delta) = \log(1-\delta) + \delta \text{ and } f''(\delta) = -\frac{1}{1-\delta} + 1.$$

Which shows that $f''(\delta) < 0$ for $\delta \in (0,1)$ and hence $f'(0) = 0$ implies that $f'(\delta) \leq 0$ in this range. Since, $f(0) = 0$ we have that $f(\delta) \leq 0$ when $\delta \in (0,1)$. Therefore,

$$\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \leq \exp\left(-\frac{\delta^2}{2}\right),$$

which completes the proof of the corollary. ■

Theorem 2.2.3 (Union Bound). *Let $E_1, \ldots, E_n$ be events. Then $\mathbb{P}\left(\bigcup_{l=1}^n E_l\right) \leq \sum_{l=1}^n \mathbb{P}(E_l)$.*

Proof. This can be proved by induction on n. When $n = 1$ the result holds clearly. Now suppose that for events $E_1, \ldots, E_k$ we have that $\mathbb{P}\left(\cup_{l=1}^k E_l\right) \leq \sum_{l=1}^k \mathbb{P}(E_l)$. Then for events $E_1, \ldots, E_k, E_{k+1}$ it follows that

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{l=1}^{k+1} E_l\right) &= \mathbb{P}\left(\bigcup_{l=1}^{k} E_l\right) + \mathbb{P}\left(E_{k+1}\right) - \mathbb{P}\left(\left(\bigcup_{l=1}^{k} E_l\right) \cap E_{k+1}\right) \\
&\leq \mathbb{P}\left(\bigcup_{l=1}^{k} E_l\right) + \mathbb{P}\left(E_{k+1}\right) \\
&\leq \sum_{l=1}^{k} \mathbb{P}\left(E_l\right) + \mathbb{P}\left(E_{k+1}\right) \\
&= \sum_{l=1}^{k+1} \mathbb{P}\left(E_l\right).
\end{aligned}$$

Therefore, by induction the result holds for all $n \in \mathbb{N}$ which completes the proof. □ ■

Proof. For the proof we consider the case when $C = 1$, with the more general case following by rescaling the loss function. For $\mathbf{w} \in \mathcal{W}$ let

$$\epsilon(\mathbf{w}) = \sqrt{\frac{2R(\mathbf{w})\left(\log\left(\frac{1}{\pi(\mathbf{w})}\right) + \log\left(\frac{1}{\delta}\right)\right)}{m}}.$$

Then Corollary 2.2.2 states that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\hat{R}(\mathbf{w}) \leq R(\mathbf{w}) - \epsilon(\mathbf{w})\right) \leq \exp\left(-\frac{m\epsilon(\mathbf{w})^2}{2R(\mathbf{w})}\right) = \delta\pi(\mathbf{w}).$$

Summing over all $\mathbf{w}$ and applying the union bound we conclude that the probability that a $\mathbf{w}$ exists with the property that $R(\mathbf{w}) > \hat{R}(\mathbf{w}) + \epsilon(\mathbf{w})$ is δ. Therefore, for all $\mathbf{w}$

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

Using $\sqrt{ab} = \inf_{\lambda>0}\left(\frac{a}{2\lambda} + \frac{\lambda b}{2}\right)$ we get that

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

which upon rearrangement completes the proof. □

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 [17] 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 [17] 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^{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 $\gamma \geq 0$ to be

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

Similarly, we have the empirical classification margin loss defined as

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

We will sometimes use $L(\cdot)$ to denote $L_0(\cdot)$ and refer to it as the classification loss. Suppose that our neural network has d fully connected layers and let x^i be the vector before the activation at layer $i = 0, \ldots, d$

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 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,

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

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 (γ, 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 $y \in \{1, \ldots, k\}$.

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 (γ, 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 $y \in \{1, \ldots, k\}$.

Theorem 2.5. *Suppose $G_{\mathcal{W},s} = \{g_{\mathbf{w},s} : \mathbf{w} \in \mathcal{W}\}$ where $\mathbf{w}$ 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 (γ, S)-compressible via $G_{\mathcal{W},s}$ with helper string s, then there exists $\mathbf{w} \in \mathcal{W}$ with high probably 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.

Theorem 2.5.1 (Hoeffding's Inequality [5])**.** *Let $X_1, \ldots, X_n$ be independent random variables with range $[a, b]$ and $\mathbb{E}(X_i) = \mu$. Then for $\epsilon > 0$ we have that*

$$(i) \quad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leq -\epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \text{ and } (ii) \quad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right) < \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Proof. Let $Z_i = X_i - \mathbb{E}(X_i)$ and $Z = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Then for $\lambda > 0$ we can apply Markov's inequality to deduce that

$$\begin{aligned}
\mathbb{P}(Z \geq \epsilon) &= \mathbb{P}\left(e^{\lambda Z} \geq e^{\lambda \epsilon}\right) \\
&\leq e^{-\lambda\epsilon}\mathbb{E}\left(e^{\lambda Z}\right) \\
&\leq e^{-\lambda\epsilon}\prod_{i=1}^{n}\mathbb{E}\left(\exp\left(\frac{\lambda Z_i}{n}\right)\right) \\
&\leq e^{-\lambda\epsilon}\prod_{i=1}^{n}\exp\left(\frac{\lambda^2(b-a)^2}{n^2}\right) \qquad \text{Lemma 2.1.3} \\
&\leq \exp\left(-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8n}\right).
\end{aligned}$$

Setting $\lambda = \frac{4n\epsilon}{(b-a)^2}$ gives

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Applying the same reasoning but for $\mathbb{P}(Z \leq -\epsilon)$ and $\lambda = -\frac{4n\epsilon}{(b-a)^2}$ give

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leq -\epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

which completes the proof of the theorem. ■

Proof. For $\mathbf{w} \in \mathcal{W}$, the empirical classification margin $\hat{L}_0(g_\mathbf{w})$ is the average of m i.i.d Bernoulli random variables with parameter $L_0(g_\mathbf{w})$. Let $X_i \sim \text{Bern}(L_0(g_\mathbf{w}))$ so that $\mu = \mathbb{E}(X_i) = L_0(g_\mathbf{w})$. It follows that

$$\begin{aligned}
\mathbb{P}\left(L_0(g_\mathbf{w}) - \hat{L}_0(g_\mathbf{w}) \geq \tau\right) &= \mathbb{P}\left(L_0(g_\mathbf{w}) - \frac{1}{m}\sum_{i=1}^{n} X_i \geq \tau\right) \\
&= \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{n} X_i - \mu \leq -\tau\right) \\
&\leq \exp\left(-2\tau^2 m\right),
\end{aligned}$$

where Hoeffding's inequality (i) has been applied. With $\tau = \sqrt{\frac{q\log(r)}{m}}$ we have that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_0(g_\mathbf{w}) \leq \hat{L}_0(g_\mathbf{w}) + \sqrt{\frac{q\log(r)}{m}}\right) \geq 1 - \exp(-2q\log(r)).$$

As there are only r^q different $\mathbf{w}$, we can apply a union bound arguments to conclude that for all $\mathbf{w} \in \mathcal{W}$ we have that

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_0(g_\mathbf{w}) \geq \hat{L}_0(g_\mathbf{w}) + \sqrt{\frac{q\log(r)}{m}}\right) &\leq r^q \exp(-q\log(r)) \\
&= \exp\left(q\log(r) - 2q\log(r)\right) \\
&= \exp\left(-q\log(r)\right).
\end{aligned}$$

Which implies that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_0(g_\mathbf{w}) \leq \hat{L}_0(g_\mathbf{w}) + \sqrt{\frac{q\log(r)}{m}}\right) \geq 1 - \exp\left(-q\log(r)\right).$$

As h is (γ, S)-compressible via $G_{\mathcal{W},S}$ then there exists a $\mathbf{w} \in \mathcal{W}$ such that for any $x \in \mathcal{X}$ and any y we have

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

Therefore, as long as h has a margin at least γ the classifier $g_\mathbf{w}$ classifies the examples correctly so that

$$\hat{L}_0(g_\mathbf{w}) \leq \hat{L}_\gamma(h).$$

Combining this with the previous observations completes the proof of the theorem. □

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 on the hypothesis class, assumptions are only made on h and its properties on a finite training set.*

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

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

Proof. The proof of this corollary proceeds in exactly the same ways as the proof of Theorem 2.5, however, in the last step we can use the upper-bound

$$\hat{L}_0(g_{\mathbf{w}}) \leq \hat{L}_\gamma(h) + \zeta.$$

Which arises as for the fraction of the training sample where the compression doesn't work we assume that the loss is maximized, which was assumed to be 1. □

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 others have label -1. The linear classifiers will be parameterized 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 $y\left(\mathbf{w}^\top x\right) \geq \gamma$ 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.

Algorithm 1 $(\gamma, \mathbf{w})$

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

Theorem 2.8. *For any number of samples m, Algorithm 1 generates a compressed vector $\hat{\mathbf{w}}$, such that*

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

Theorem 2.8.1 (Chebyshev's Inequality)**.** *For a random variable X, with with variance $\sigma^2 \in (0, \infty)$ and mean $\mu < \infty$, then for $k > 0$ we have that*

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. To prove Chebyshev's inequality we use Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(|X - \mu| \geq k\sigma\right) &= \mathbb{P}\left(|X - \mu|^2 \geq k^2\sigma^2\right) \\
&\leq \frac{\mathbb{E}\left((X - \mu)^2\right)}{k^2\sigma^2} \\
&= \frac{\sigma^2}{k^2\sigma^2} \\
&= \frac{1}{k^2}
\end{aligned}$$

which completes the proof of the theorem. ■

Theorem 2.8.2 (Relative Chernoff Bound 2 [5])**.** *Suppose $X_1, \ldots, X_n$ are independent random variables with range $\{0, 1\}$. Let $\mu = \sum_{i=1}^n X_i$. Then for $\delta > 0$ we have*

$$\mathbb{P}\left(X \geq (1 + \delta)\mu\right) \leq \left(\frac{e^\delta}{(1 + \delta)^{(1+\delta)}}\right)^\mu.$$

Proof. Using Markov's inequality we note that for $t > 0$ we have

$$\begin{aligned}\mathbb{P}\left(X \geq (1+\delta)\mu\right) &= \mathbb{P}\left(e^{tX} \geq e^{t(1+\delta)\mu}\right) \\ &\leq \frac{\mathbb{E}\left(e^{tX}\right)}{e^{t(1+\delta)\mu}} \\ &\leq \frac{\exp\left(\left(e^t-1\right)\mu\right)}{\exp\left(t(1+\delta)\mu\right)}.\end{aligned}$$

Setting $t = \log(1+\delta) > 0$ we get that

$$\mathbb{P}\left(X \geq (1+\delta)\mu\right) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu.$$

which completes the proof of the theorem. ■

Corollary 2.8.3 ([5])**.** *Suppose $X_1, \ldots, X_n$ are independent random variables with range $\{0,1\}$. Let $\mu = \sum_{i=1}^n X_i$. Then for $\delta \in (0,1]$ we have*

$$\mathbb{P}\left(X \geq (1+\delta)\mu\right) \leq \exp\left(-\frac{\mu\delta^2}{3}\right).$$

Proof. Consider

$$f(\delta) = \delta - (1+\delta)\log(1+\delta) + \frac{\delta^2}{3}$$

for $\delta \in (0,1]$. Note that

$$f'(\delta) = -\log(1+\delta) + \frac{2}{3}\delta \text{ and } f''(\delta) = -\frac{1}{1+\delta} + \frac{2}{3}.$$

Which shows that $f''(\delta) < 0$ for $\delta \in \left[0, \frac{1}{2}\right)$ and $f''(\delta) < 0$ for $\delta > \frac{1}{2}$. Since $f'(0) = 0$ and $f'(1) < 0$ we deduce that $f'(\delta) \leq 0$ in the interval $[0,1]$. Since, $f(0) = 0$ we have that $f(\delta) \leq 0$ when $\delta \in [0,1]$. Therefore,

$$\frac{e^\delta}{(1+\delta)^{1+\delta}} \leq \exp\left(-\frac{\delta^2}{3}\right),$$

which completes the proof of the corollary. ■

Lemma 2.8.4. *Algorithm 1 $(\gamma, \mathbf{w})$ returns a vector $\hat{\mathbf{w}}$ such that for any fixed u, with probability $1-\eta$, $\left|\hat{\mathbf{w}}^\top u - \mathbf{w}^\top u\right| \leq \gamma$. The vector $\hat{\mathbf{w}}$ has at most $O\left(\frac{\log d}{\eta\gamma^2}\right)$ non-zero entries with high probability.*

Proof. By the construction of Algorithm 1 it is clear that for all i we have $\mathbb{E}\left(\hat{w}_i\right) = w_i$. Similarly, we have that

$$\mathrm{Var}\left(\hat{w}_i\right) = 2p_i(1-p_i)\frac{w_i^2}{p_i^2} \leq \frac{2w_i^2}{p_i} = \eta\gamma^2.$$

Therefore, for u independent of $\hat{\mathbf{w}}$ we have that

$$\mathbb{E}\left(\hat{\mathbf{w}}^\top u\right) = \sum_{i=1}^d \mathbb{E}\left(\hat{w}_i u_i\right) = \sum_{i=1}^d \mathbb{E}\left(\hat{w}_i\right) u_i = \sum_{i=1}^d w_i u_i = \mathbf{w}^\top u,$$

and

$$\mathrm{Var}\left(\hat{\mathbf{w}}u^\top\right) = \mathrm{Var}\left(\sum_{i=1}^{d}\hat{w}_i u_i\right) = \sum_{i=1}^{d}\mathrm{Var}(w_i)u_i^2 \leq \eta\gamma^2\sum_{i=1}^{d}u_i^2 = \eta\gamma^2\|u\|^2 \leq \eta\gamma^2.$$

Therefore, by Chebyshev's inequality we have that

$$\mathbb{P}\left(\left|\hat{\mathbf{w}}^\top u - \mathbf{w}^\top u\right| \geq \gamma\right) \leq \eta.$$

To determine how we can bound the number of non-zero entries we analyze the behavior of the right-hand side of Theorem 2.8.2. For each entry we can define the indicator random variable X_i which is 1 when the entry is non-zero and 0 otherwise. Note that $\mathbb{E}(X_i) = p_i$ and for $X = \sum_{i=1}^{d} X_i$ we have that

$$\mu = \mathbb{E}(X) = \sum_{i=1}^{d} p_i = \frac{2\|\mathbf{w}\|^2}{\eta\gamma^2} \leq \frac{2}{\eta\gamma^2}.$$

Therefore, we need to find for what order function $f(\cdot)$ does

$$\frac{e^{f(d)-1}}{f(d)^{f(d)}} \to 0, \quad \text{as } d \to \infty,$$

so that the number of non-zero elements is bounded by $O\left(\frac{f(d)}{\eta\gamma^2}\right)$ with high probability using Theorem 2.2.1. We observe that with $f(d) = \log(d)$ we get the desired convergence, and so this completes the proof of the lemma. ■

In the discrete case, a similar result holds. For a vector $\mathbf{w} \in \mathbb{R}^d$ and for a given pair (η, γ) let its discrete version be $\hat{\mathbf{w}}$ where

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

Let its capped version be $\mathbf{w}^*$ where

$$w_i^* = \begin{cases} 0 & |\tilde{w}_i| \geq 2\eta\gamma\sqrt{d} \\ w_i & \text{Otherwise.} \end{cases}$$

Let its truncated version be $\mathbf{w}'$ where

$$w_i' = \begin{cases} w_i & |w_i| \geq \frac{\gamma}{4\sqrt{d}} \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 2.8.5. *Let Algorithm 1 $\left(\frac{\gamma}{2}, \mathbf{w}\right)$ return the vector $\tilde{\mathbf{w}}$. Then for any fixed u with probability at least $1-\eta$, we have that*

$$\left|\hat{\tilde{\mathbf{w}}}^\top u - \mathbf{w}^\top u\right| \leq \gamma.$$

Proof. First note that

$$\|\mathbf{w}' - \mathbf{w}\|^2 = \sum_{i=1}^{d}|w_i' - w_i|^2 \leq \sum_{i=1}^{d}\frac{\gamma^2}{16d} = \frac{\gamma^2}{16},$$

which implies that $\|\mathbf{w}' - \mathbf{w}\| \leq \frac{\gamma}{4}$. Similarly,

$$\left\|\hat{\tilde{\mathbf{w}}} - \tilde{\mathbf{w}}^*\right\|^2 = \sum_{i=1}^{d}\left|\hat{\tilde{w}}_i - \tilde{w}_i^*\right|^2 \leq \sum_{i=1}^{d}\left(\frac{1}{2}\frac{\gamma}{2\sqrt{d}}\right)^2 = \frac{\gamma^2}{16},$$

which implies that $\left\|\hat{\tilde{\mathbf{w}}} - \tilde{\mathbf{w}}^*\right\| \leq \frac{\gamma}{4}$. Now suppose Algorithm 2 $\left(\frac{\gamma}{2}, \mathbf{w}'\right)$ returns the capped vector $\mathbf{v}$.

When $|w_i| \leq \frac{\gamma}{4\sqrt{d}}$ we have that $w_i' = 0$ so that $v_i = 0$. We also have that, $\tilde{w}_i$ is either 0 or

$$|\tilde{w}_i| = \left| \frac{w_i}{\left(\frac{2w_i^2}{\eta\gamma^2}\right)} \right| = \left| \frac{\eta\gamma^2}{2w_i} \right| \geq |2\eta\gamma\sqrt{d}|$$

so that in any case we also have that $\tilde{w}_i^* = 0$. If instead $|\tilde{w}_i| \geq 2\eta\gamma\sqrt{d}$ then $\tilde{w}_i^* = 0$ and through similar computations we have that $|w_i| \leq \frac{\gamma}{4\sqrt{d}}$ and so $v_i = 0$. It is clear that when either of these two conditions do not hold we have $\tilde{w}_i^* = v_i$ as $w_i = w_i'$. Therefore, $\tilde{\mathbf{w}}^* = \mathbf{v}$ and so from Lemma 2.8.4 we conclude that with probability at least $1 - \eta$ we have $\left|(\tilde{\mathbf{w}}^*)^\top u - (\mathbf{w}')^\top u\right| \leq \frac{\gamma}{2}$ for a fixed vector u with $\|u\| \leq 1$. Using these observation we deduce for a fixed vector u with $\|u\| \leq 1$ that

$$\begin{aligned}
\left|\hat{\tilde{\mathbf{w}}}^\top u - \mathbf{w}^\top u\right| &\leq \left|\hat{\tilde{\mathbf{w}}}^\top u - (\tilde{\mathbf{w}}^*)^\top u\right| + \left|(\hat{\mathbf{w}}^*)^\top u - (\mathbf{w}')^\top u\right| + \left|(\mathbf{w}')^\top u - \mathbf{w}^\top u\right| \\
&\leq \left\|\hat{\tilde{\mathbf{w}}}' - \tilde{\mathbf{w}}^*\right\| + \frac{\gamma}{2} + \|\mathbf{w}' - \mathbf{w}\| \\
&\leq \frac{\gamma}{4} + \frac{\gamma}{2} + \frac{\gamma}{4} = \gamma,
\end{aligned}$$

with probability at least $1 - \eta$, which completes the proof of the lemma. ■

Proof. Now choose $\eta = \left(\frac{1}{\gamma^2 m}\right)^{\frac{1}{3}}$. By Lemma 2.8.4 and Lemma 2.8.5 we know that Algorithm 1 works with probability $1-\eta$ and has at most $\tilde{O}\left(\frac{\log(d)}{\eta\gamma^2}\right)$ parameters each of which can take some finite number r of discrete values. Using Corollary 2.7 we know that

$$L\left(\hat{\mathbf{w}}\right) \leq O\left(\eta + \sqrt{\frac{\log(d)\log(r)}{\eta\gamma^2 m}}\right) \leq \tilde{O}\left(\eta + \sqrt{\frac{1}{\eta\gamma^2 m}}\right) \leq \tilde{O}\left(\left(\frac{1}{\gamma^2 m}\right)^{\frac{1}{3}}\right)$$

which completes the proof of the theorem. □

Remark 2.9. *The rate is not optimal as it depends on $m^{1/3}$ and not $\sqrt{m}$. To resolve this we employ helper strings.*

Algorithm 2 $(\gamma, \mathbf{w})$

Require: vector $\mathbf{w}$ with $\|\mathbf{w}\| \leq 1$, η.
Ensure: vector $\hat{\mathbf{w}}$ such that for any fixed vector $\|u\| \leq 1$, with probability at least $1 - \eta$, $|\mathbf{w}^\top u - \hat{\mathbf{w}}^\top u| \leq \gamma$.

Let $k = \frac{16\log\left(\frac{1}{\eta}\right)}{\gamma^2}$.
Sample the random vectors $v_1, \ldots, v_k \sim \mathcal{N}(0, I)$.
Let $z_i = \langle v_i, \mathbf{w}\rangle$.
(In Discrete Case) Round z_i to closes multiple of $\frac{\gamma}{2\sqrt{dk}}$.
return $\hat{\mathbf{w}} = \frac{1}{k}\sum_{i=1}^k z_i v_i$

Remark 2.10. *The vectors v_i of Algorithm 2 form the helper string.*

Theorem 2.11. *For any number of sample m, Algorithm 2 with the helper string generates a compressed vector $\hat{\mathbf{w}}$, such that*

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

Lemma 2.11.1. *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 $\left|\hat{\mathbf{w}}^\top u - \mathbf{w}^\top u\right| \leq \gamma$.*

Proof. Observe that

$$\hat{\mathbf{w}}^\top u = \frac{1}{k}\sum_{i=1}^{k}\langle v_i, \mathbf{w}\rangle\langle v_i, u\rangle.$$

Where,

$$\mathbb{E}\left(\langle v_i, \mathbf{w}\rangle\langle v_i, u\rangle\right) = \mathbb{E}\left(\mathbf{w}^\top v_i v_i^\top u\right) = \mathbf{w}^\top \mathbb{E}\left(v_i v_i^\top\right) = \mathbf{w}^\top u$$

and

$$\mathrm{Var}\left(\hat{\mathbf{w}}^\top u\right) \leq O\left(\frac{1}{k}\right).$$

Therefore, by standard concentration inequalities

$$\mathbb{P}\left(\left|\hat{\mathbf{w}}^\top u - \mathbf{w}^\top u\right| \geq \frac{\gamma}{2}\right) \leq \exp\left(\frac{-\gamma^2 k}{16}\right) \leq \eta.$$

As with discretization the vector can only change by at most $\frac{\gamma}{2}$, the proof of the lemma is complete. ■

Proof. Choosing $\eta = \frac{1}{m}$ and applying Lemma 2.11.1 we see that with probability $1-\eta$, the compressed vector has at most

$$O\left(\frac{\log(m)}{\gamma^2}\right)$$

parameters. As the number of parameters is finite we can assume that there is a finite number of discrete values, r, that each parameter can take. For example, if M is the large absolute value of the parameter then we can take $r = 2\frac{M}{\left(\frac{\gamma}{2\sqrt{dk}}\right)} + 1$. Therefore, from Corollary 2.7 we know that

$$L\left(\mathbf{w}\right) \leq O\left(\eta + \sqrt{\frac{\frac{\log(m)}{\gamma^2}\log(r)}{m}}\right) \leq \tilde{O}\left(\sqrt{\frac{1}{\gamma^2 m}}\right)$$

which completes the proof of the theorem. □

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 algorithm 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.

Definition 2.12. *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 x with respect to $\mathcal{N}$ is*

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

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

Remark 2.13. *When $x \neq 0$ and the noise distribution is the Gaussian distribution $\mathcal{N}(0, I)$ then the noise sensitivity of matrix M is exactly $\frac{\|M\|_F^2}{\|Mx\|^2}$. Hence, it is at most the stable rank of M.*

Algorithm 3 (A, ϵ, η)

Require: Layer matrix $A \in \mathbb{R}^{n_1 \times n_2}$, error parameters ϵ, η.
Ensure: Returns $\hat{A}$ such that for all vectors u, v we have that

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

Sample $k = \frac{\log(\frac{1}{\eta})}{\epsilon^2}$ random matrices $M_1, \ldots, M_k$ with i.i.d entries ± 1.
for $k' = 1 \to k$ **do**
  Let $Z_l = \langle A, M_l \rangle M_l$
end for
return $\hat{A} = \frac{1}{k} \sum_{l=1}^{k} Z_l$

Definition 2.14. *The layer cushion of layer i is defined as the largest μ_i such that for any $x \in \mathcal{X}$ we have*

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

Remark 2.15. *Note that $\frac{1}{\mu_i^2}$ is equal to the noise sensitivity of A^i at $\phi\left(x^{i-1}\right)$ with respect to noise $\eta \sim \mathcal{N}(0, I)$.*

Definition 2.16. *For layers $i \leq j$ the inter-layer cushion $\mu_{i,j}$ is the largest number such that*

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

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

Remark 2.17. *Note that $J_{x^i}^{i,i} = I$ so that*

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

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 2.18. *The activation contraction c is defined as the smallest number such that for any layer i*

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

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

Definition 2.19. *Let η be the noise generated as a result of applying Algorithm 3 to some of the layers before layer i. Define the inter-layer smoothness ρ_δ to be the smallest number such that with probability $1 - \delta$ and for layers $i < j$ we have that*

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

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

Remark 2.20. *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 $U\hat{A}x$.*

Theorem 2.21. *For any fully connected network $h_{\mathbf{w}}$ with $\rho_\delta \geq 3d$, any probability $0 < \delta \leq 1$ and any margin γ. Algorithm 3 generates weights $\tilde{\mathbf{w}}$ such that*

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

Lemma 2.21.1. *For any $0 < \delta$ and $\epsilon \leq 1$ let $G = \left\{\left(U^i, x^i\right)\right\}_{i=1}^m$ 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 $\left(A, \epsilon, \eta = \frac{\delta}{mk}\right)$. With probability at least $1 - \delta$ we have for any $(U, x) \in G$ that $\left\|U(\hat{A} - A)x\right\| \leq \epsilon \|A\|_F \|U\|_F \|x\|$.*

Proof. For fixed vectors u, v we have that

$$u^\top \hat{A} v = \frac{1}{k} \sum_{l=1}^{k} u^\top Z_l v = \frac{1}{k} \sum_{l=1}^{k} \langle A, M_l \rangle \left\langle uv^\top, M_l \right\rangle.$$

Furthermore, we have that

$$\begin{aligned}
\mathbb{E}\left(\langle A, M_l \rangle \left\langle uv^\top, M_l \right\rangle\right) &= \mathbb{E}\left(\operatorname{tr}\left(A^\top M_l\right) \operatorname{tr}\left(M_l^\top \left(uv^\top\right)\right)\right) \\
&= \mathbb{E}\left(\left(\sum_{i,j} a_{ij} m_{ij}\right)\left(\sum_{i,j} m_{ij} u_{ij}\right)\right) \\
&= \sum_{i,j} a_{ij} u_{ij} \\
&= \left\langle A, uv^\top \right\rangle,
\end{aligned}$$

where we achieve the last equality as we note that $\mathbb{E}(m_{ij} m_{kl})$ is 1 when $ij = kl$ and 0 otherwise. By standard concentration inequalities we deduce that

$$\mathbb{P}\left(\left|\frac{1}{k} \sum_{l=1}^{k} \langle A, M_l \rangle \left\langle uv^\top, M_l \right\rangle - \left\langle A, uv^\top \right\rangle\right| \geq \epsilon \|A\|_F \left\|uv^\top\right\|_F\right) \leq \exp\left(-k\epsilon^2\right).$$

Therefore, for the choice of k from Algorithm 3 we know that

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

Let $(U, x) \in G$ and u_i be the i^{th} row of U. We can apply the above result with a union bound over the mn rows u_i in G to get that

$$\mathbb{P}\left(\left|u_i^\top \hat{A} v - u_i^\top A v\right| \leq \epsilon \|A\|_F \|u_i\| \|v\|\right) \geq 1 - \delta$$

for all i simultaneously. Furthermore,

$$\left\|U\left(\hat{A} - A\right)x\right\|^2 = \sum_{i=1}^{n} \left(u_i^\top \left(\hat{A} - A\right) x\right)^2, \text{ and } \|U\|_F^2 = \sum_{i=1}^{n} \|u_i\|^2$$

we see that with probability at least $1 - \delta$ we have

$$\begin{aligned}
\left\|U(\hat{A} - A)x\right\|^2 &= \sum_{i=1}^{n} \left(u_i^\top \left(\hat{A} - A\right) x\right)^2 \\
&\leq \sum_{i=1}^{n} \epsilon^2 \|A\|_F^2 \|u_i\|^2 \|x\| \\
&= \epsilon^2 \|A\|_F^2 \|U\|^2 \|x\|
\end{aligned}$$

which completes the proof of the lemma. ■

Lemma 2.21.2. *For any fully connected network $h_{\mathbf{w}}$ with $\rho_\delta \geq 3d$, any probability $0 < \delta \leq 1$ and any $0 < \epsilon \leq 1$, Algorithm 3 can generate weights $\tilde{\mathbf{w}}$ for a network with at most*

$$\frac{32c^2d^2\log\left(\frac{mdn}{\delta}\right)}{\epsilon^2}\cdot\sum_{i=1}^{d}\frac{1}{\mu_i^2\mu_{i\to}^2}$$

total parameters such that for any $x \in \mathcal{X}$ we have

$$\mathbb{P}\left(\|h_{\mathbf{w}}(x) - h_{\tilde{\mathbf{w}}}(x)\| \leq \epsilon\|h_{\mathbf{w}}(x)\|\right) \geq 1 - \frac{\delta}{2},$$

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

Proof. The proof of this lemma proceeds by induction. For $i \geq 0$ let $\hat{x}_i^j$ be the output at layer j if weights $A^1, \ldots, A^i$ are replaced with $\tilde{A}^1, \ldots, \tilde{A}^i$. We want to show for any i if $j \geq i$ then

$$\mathbb{P}\left(\left\|\hat{x}_i^j - x^j\right\| \leq \frac{i}{d}\epsilon\left\|x^j\right\|\right) \geq 1 - \frac{i\delta}{2d}.$$

For $i = 0$ the result is clear as the weight matrices are unchanged. Suppose the result holds true for $i-1$. Let $\hat{A}^i$ be the result of applying Algorithm 3 to A^i with $\epsilon_i = \frac{\epsilon\mu_i\mu_{i\to}}{4cd}$ and $\eta = \frac{\delta}{6d^2n^2m}$. Consider the set

$$G = \left\{\left(J_{x^i}^{i,j}, x^i\right) : x^i \in \mathcal{X}, j \geq i\right\}$$

and let $\Delta^i = \hat{A}^i - A^i$. Note that

$$\left\|\hat{x}_i^j - x^j\right\| \leq \left\|\hat{x}_i^j - \hat{x}_{i-1}^j\right\| + \left\|\hat{x}_{i-1}^j - x^j\right\|.$$

The second term is bounded by $\frac{(i-1)\epsilon\left\|x^j\right\|}{d}$ by inductive assumption. Therefore, it suffices to show that the first term is bounded by $\frac{\epsilon}{d}$ to complete the inductive step. First observe that,

$$\left\|\hat{x}_i^j - \hat{x}_{i-1}^j\right\| \leq \left\|J_{x^i}^{i,j}\left(\Delta^i\phi\left(\hat{x}^{i-1}\right)\right)\right\| + \left\|M^{i,j}\left(\hat{A}^i\phi\left(\hat{x}^{i-1}\right)\right) - M^{i,j}\left(A^i\phi\left(\hat{x}^{i-1}\right)\right) - J_{x^i}^{i,j}\left(\Delta^i\phi\left(\hat{x}^{i-1}\right)\right)\right\|.$$

The first term can be bounded as follows

$$\begin{aligned}
\left\|J_{x^i}^{i,j}\left(\Delta^i\phi\left(\hat{x}^{i-1}\right)\right)\right\| &\leq \frac{\epsilon\mu_i\mu_{i\to}}{6cd}\left\|J_{x^i}^{i,j}\right\|\left\|A^i\right\|_F\left\|\phi\left(\hat{x}^{i-1}\right)\right\| \quad \text{Lemma 2.21.1}\\
&\leq \frac{\epsilon\mu_i\mu_{i\to}}{6cd}\left\|J_{x^i}^{i,j}\right\|\left\|A^i\right\|_F\left\|\hat{x}^{i-1}\right\| \quad \text{Lipschitz of } \phi\\
&\leq \frac{\epsilon\mu_i\mu_{i\to}}{3cd}\left\|J_{x^i}^{i,j}\right\|\left\|A^i\right\|_F\left\|x^{i-1}\right\| \quad \text{Inductive Assumption}\\
&\leq \frac{\epsilon\mu_{i\to}}{3d}\left\|J_{x^i}^{i,j}\right\|\left\|A^i\phi\left(x^{i-1}\right)\right\|\\
&\leq \frac{\epsilon\mu_{i\to}}{3d}\left\|J_{x^i}^{i,j}\right\|\left\|x^i\right\|\\
&\leq \frac{\epsilon}{3d}\left\|x^j\right\|.
\end{aligned}$$

The second term can be split as

$$\left\|\left(M^{i,j} - J_{x^i}^{i,j}\right)\left(\hat{A}^i\phi\left(\hat{x}^{i-1}\right)\right)\right\| + \left\|\left(M^{i,j} - J_{x^i}^{i,j}\right)\left(A^i\phi\left(\hat{x}^{i-1}\right)\right)\right\|,$$

which can be bounded by inter-layer smoothness. By inductive assumption

$$\left\|A^i\phi\left(\hat{x}^{i-1}\right)-x^i\right\| \leq \frac{(a-1)\epsilon\left\|x^i\right\|}{d} \leq \epsilon\left\|x^i\right\|.$$

Then by inter-layer smoothness

$$\left\|\left(M^{i,j}-J^{i,j}_{x^i}\right)\left(A^i\phi\left(\hat{x}^{i-1}\right)\right)\right\| \leq \frac{\left\|x^b\right\|\epsilon}{\rho_\delta} \leq \frac{\epsilon}{3d}\left\|x^j\right\|.$$

On the other hand,

$$\left\|\hat{A}^i\phi\left(\hat{x}^{i-1}\right)-x^i\right\| \leq \left\|A^i\phi\left(\hat{x}^{i-1}\right)-x^i\right\| + \left\|\Delta^i\phi\left(\hat{x}^{i-1}\right)\right\| \leq \frac{(i-1)\epsilon}{d}+\frac{\epsilon}{3d} \leq \epsilon$$

so that

$$\left\|\left(M^{i,j}-J^{i,j}_{x^i}\right)\left(A^i\phi\left(\hat{x}^{i-1}\right)\right)\right\| \leq \frac{\epsilon}{3d}\left\|x^j\right\|.$$

This completes the inductive step. To complete the proof we bound the total number of parameters. With the values of ϵ_i and η used in the induction we see that the total number of parameters is

$$\begin{aligned}\sum_{i=1}^{d}\frac{\log\left(\frac{1}{\frac{\delta}{6d^2n^2m}}\right)}{\left(\frac{\epsilon\mu_i\mu_{i\to}}{4cd}\right)^2} &= \sum_{i=1}^{d}\frac{16c^2d^2\log\left(\frac{6d^2n^2m}{\delta}\right)}{\epsilon^2\mu_i\mu_{i\to}}\\ &\leq \frac{16c^2d^2\log\left(\frac{d^2n^2m^2}{\delta^2}\right)}{\epsilon^2}\sum_{i=1}^{d}\frac{1}{\mu_i^2\mu_{i\to}^2}\\ &\leq \frac{32c^2d^2\log\left(\frac{mdn}{\delta}\right)}{\epsilon^2}\sum_{i=1}^{d}\frac{1}{\mu_i^2\mu_{i\to}^2},\end{aligned}$$

which completes the proof of the lemma. ■

Lemma 2.21.3. *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 $h_{\mathbf{w}}$ 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*

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

Proof. In the first case suppose that $\gamma^2 > 2\max_{x\in\mathcal{X}}\|h_{\mathbf{w}}(x)\|_2^2$, then

$$\left|h_{\mathbf{w}}(x)[y]-\max_{j\neq y}h_{\mathbf{w}}(x)[j]\right|^2 \leq 2\max_{x\in\mathcal{X}}\|h_{\mathbf{w}}(x)\|_2^2 \leq \gamma^2.$$

Therefore, the margin can be at most γ which implies that $\hat{L}_\gamma(h_{\mathbf{w}})=1$ and so the statement holds in this case. If instead $\gamma^2 \leq 2\max_{x\in\mathcal{X}}\|h_{\mathbf{w}}(x)\|_2^2$, then setting $\epsilon=\frac{\gamma^2}{2\max_{x\in\mathcal{X}}\|h_{\mathbf{w}}(x)\|_2^2}$ in Lemma 2.21.2 we conclude that

$$\|h_{\mathbf{w}}(x)-h_{\tilde{\mathbf{w}}}(x)\| \leq \frac{\gamma}{\sqrt{2}}$$

for any $x\in\mathcal{X}$. If for $(x,y)\in\mathcal{Z}$ we have that

$$h_{\mathbf{w}}(x)[y] \leq \gamma+\max_{j\neq y}f_{\mathbf{w}}[j]$$

then for $h_{\tilde{\mathbf{w}}}$ the contribution of this data point to the empirical classification loss can only be less than or equal to its contribution to the empirical classification margin loss of $h_{\mathbf{w}}$. On the other hand, suppose that

$$h_{\mathbf{w}}[y] > \gamma + \max_{j \neq y} h_{\mathbf{w}}[j],$$

so that the data point doesn't contribute to $\hat{L}_\gamma(h_{\mathbf{w}})$. For the data point's contribution to $\hat{L}_0(h_{\tilde{\mathbf{w}}})$ to be larger we need

$$h_{\tilde{w}} \leq \max_{j \neq y} h_{\tilde{w}}(x)[y]$$

which would either require a change of more than γ between two components of $h_{\mathbf{w}}(x)$ under compression to $h_{\tilde{\mathbf{w}}}(x)$. If this change occurs then the distance between $h_{\mathbf{w}}(x)$ and $h_{\tilde{\mathbf{w}}}(x)$ is minimized when only two components change by more than $\frac{\gamma}{2}$. Suppose that components the i^{th} and j^{th} components move sufficiently then,

$$\begin{aligned}
\|h_{\mathbf{w}}(x) - h_{\tilde{w}}(x)\|^2 &> |h_{\mathbf{w}}(x)[i] - h_{\tilde{\mathbf{w}}}(x)[i]|^2 + |h_{\mathbf{w}}(x)[j] - h_{\tilde{\mathbf{w}}}(x)[j]|^2 \\
&> \frac{\gamma^2}{4} + \frac{\gamma^2}{4} \\
&= \frac{\gamma^2}{2}
\end{aligned}$$

which contradicts the conclusion of Lemma 2.21.2. We now bound the number of total parameters. With our value for ϵ Lemma 2.21.2 tells us that the number of total parameters is at most

$$\frac{32c^2d^2\log\left(\frac{mdn}{\delta}\right)}{\frac{\gamma^2}{2\max_{x\in\mathcal{X}}}} \sum_{i=1}^{d} \frac{1}{\mu_i^2 \mu_{i\to}^2},$$

from which the required result follows and completes the proof of the lemma. ■

Lemma 2.21.4. *For any matrix A let $\hat{A}$ be the truncated version of A where singular values that are smaller than $\delta \|A\|_2$.Let $h_{\mathbf{w}}$ be a d-layer network with weights $A = \{A^1, \ldots, A^d\}$. Then for any input x, weights A and $\hat{A}$, if for any layer i, $\left\|A^i - \hat{A}^i\right\| \leq \frac{1}{d}\|A^i\|$, then*

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

Proof ([16]). For $x \in \mathcal{X}$ recall that x^i is the vector before activation at layer i. Now we also consider $\hat{x}^i$ as the vector before activation for the network with weights $\hat{A}$. For the lemma $x \in \mathcal{X}$ is fixed so let $\xi_i = \left|\hat{x}^i - x^i\right|$. Now proceed by induction on the statement

$$\xi_i \leq \left(1 + \frac{1}{d}\right)^i \|x\| \left(\prod_{j=1}^{i} \left\|A^j\right\|_2\right) \sum_{j=1}^{i} \frac{\left\|A^j - \hat{A}^j\right\|_2}{\|A^j\|_2}.$$

The base case clearly holds as $\xi_0 = 0$. Therefore, for $i \geq 1$ we proceed as follows,

$$\begin{aligned}
\xi_{i+1} &= \left|\hat{A}^{i+1}\phi\left(\hat{x}^i\right) - A^{i+1}\phi\left(x^i\right)\right|_2 \\
&= \left|\hat{A}^i\left(\phi\left(\hat{x}^i\right) - \phi\left(x^i\right)\right) + \left(\hat{A}^i - A^i\right)\phi\left(x^i\right)\right|_2 \\
&\leq \left(\left\|A^{i+1}\right\|_2 + \left\|\hat{A}^{i+1} + A^{i+1}\right\|_2\right)\left|\phi\left(\hat{x}^i\right) - \phi\left(x^i\right)\right|_2 + \left\|\hat{A}^{i+1} - A^{i+1}\right\|_2\left|\phi\left(x^i\right)\right|_2 \\
&\leq \left(\left\|A^{i+1}\right\|_2 + \left\|\hat{A}^{i+1} + A^{i+1}\right\|_2\right)\left|\hat{x}^i - x^i\right|_2 + \left\|\hat{A}^{i+1} - A^{i+1}\right\|_2\left|x^i\right|_2 \\
&= \left(\left\|A^{i+1}\right\|_2 + \left\|\hat{A}^{i+1} + A^{i+1}\right\|_2\right)\xi_i + \left\|\hat{A}^{i+1} - A^{i+1}\right\|_2\left|x^i\right|_2,
\end{aligned}$$

note how the second arises as a specific property of the ReLU activation function. By the assumption of the lemma and the inductive assumption it therefore follows that

$$\begin{aligned}
\xi_{i+1} &\leq \xi_i\left(1 + \frac{1}{d}\right)\left\|A^{i+1}\right\|_2 + \left\|\hat{A}^{i+1} - A^{i+1}\right\|_2\|x\|_2\prod_{j=1}^{i}\left\|A^j\right\|_2 \\
&\leq \left(1 + \frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\left\|A^j\right\|_2\right)\|x\|_2\sum_{j=1}^{i}\frac{\left\|\hat{A}^j - A^j\right\|_2}{\left\|A^j\right\|_2} + \frac{\left\|\hat{A}^{i+1} - A^{i+1}\right\|_2}{\left\|A^{i+1}\right\|_2}\|x\|_2\prod_{j=1}^{i+1}\left\|A^j\right\|_2 \\
&\leq \left(1 + \frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\left\|A^j\right\|_2\right)\|x\|_2\sum_{j=1}^{i+1}\frac{\left\|\hat{A}^j - A^j\right\|_2}{\left\|A^j\right\|_2},
\end{aligned}$$

then using the fact that $\left(1 + \frac{1}{d}\right)^d \leq e$ completes the proof of the lemma. ■

Proof. We can assume without loss of generality that for any $i \neq j$ that

$$\|A_i\|_F = \|A_j\|_F = \beta.$$

This is because we can re-balance the matrices if necessary without effecting the cushion of the network. Therefore, for any $x \in \mathcal{X}$ we have

$$\beta^d = \prod_{i=1}^{d}\left\|A^i\right\|_F \leq \frac{c\left\|x^1\right\|}{\|x\|\mu_1}\prod_{i=2}^{d}\left\|A^i\right\|_F \leq \frac{c\left\|x^2\right\|}{\|x\|\mu_1\mu_2}\prod_{i=3}^{d}\left\|A^i\right\|_F \leq \cdots \leq \frac{c^d\left\|h_{\mathbf{w}}(x)\right\|}{\|x\|\prod_{i=1}^{d}\mu_i}.$$

Now for each layer matrix A^i we can define the single layer network by $A^i x^{i-1} = x^{i-1}$ and so with $\epsilon = \min\left(\frac{1}{d}, \frac{\gamma^2}{2\max\|A^i x^{i-1}\|_2^2}\right)$ we can deduce from Lemma 2.21.2 that with probability at least $1 - \frac{\delta}{2}$ we have that

$$\left\|\tilde{A}^i - A^i\right\|_F \leq \frac{1}{d}\left\|A^i\right\|.$$

Which implies that $\left\|\tilde{A}^i\right\| \leq \beta\left(1 + \frac{1}{d}\right)$. Next we that as

$$\tilde{A}^i = \frac{1}{k}\sum_{l=1}^{k}\left\langle A^i, M_l\right\rangle M_l,$$

if $\hat{A}^i$ are the approximations of $\tilde{A}^i$ with accuracy μ then

$$\left\|\hat{A}^i - \tilde{A}^i\right\|_F \leq \sqrt{k}h\nu \leq \sqrt{q}h\nu$$

where q is the total number of parameters. Therefore, by Lemma 2.21.4 we have that

$$\begin{aligned}
|l_\gamma(h_{\tilde{w}}(x), y) - l_\gamma(h_{\hat{\mathbf{w}}}(x), y)| &\leq \frac{2e}{\gamma}\|x\| \left(\prod_{i=1}^{d} \left\|\tilde{A}^i\right\|\right) \sum_{i=1}^{d} \frac{\left\|\tilde{A}^i - \hat{A}^i\right\|_F}{\left\|\tilde{A}^i\right\|_F} \\
&\leq \frac{e^2}{\gamma}\|x\|\beta^{d-1} \sum_{i=1}^{d} \left\|\tilde{A}^i - \hat{A}^i\right\|_F \\
&\leq \frac{e^2 c^d \left\|h_{\mathbf{w}}(x)\right\| \sum_{i=1}^{d} \left\|\tilde{A}^i - \hat{A}^i\right\|_F}{\gamma\beta \prod_{i=1}^{d} \mu_i} \\
&\leq \frac{qn\nu}{\beta}, \quad \text{By Lemma 2.21.3 : } \frac{e^2 d \left\|h_{\mathbf{w}}(x)\right\|}{\gamma\beta \prod_{i=1}^{d} \mu_i} \leq \sqrt{q}.
\end{aligned}$$

From Algorithm 3 we see that

$$\beta = \left\|\hat{A}^i\right\|_F = \frac{1}{k}\sum_{l=1}^{k} \|Z_l\|_F \leq \frac{1}{k}\sum_{l=1}^{k} n|\langle A, M_l\rangle|$$

and so using $k \leq n^2$ we see that parameter $\langle A, M_l\rangle$ has absolute value at most βn. Therefore, to get an ϵ-cover with r choices for each parameter we require that

$$\frac{qn\left(\frac{\beta n}{r}\right)}{\beta} = \frac{qn^2}{r} \leq \epsilon$$

which means that $r \geq \frac{qn^2}{\epsilon}$. We need these choices over q parameters and so the covering number is given by

$$\left(\frac{q\beta n^2}{\epsilon}\right)^q \leq \left(\frac{q\beta n}{\epsilon}\right)^{2q}.$$

Therefore, we can bound the Rademacher complexity by

$$\epsilon + \sqrt{\frac{4q\log\left(\frac{q\beta n}{\epsilon}\right)}{m}}$$

from which we can deduce that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_0\left(h_{\mathbf{w}}\right) \leq \hat{L}_0\left(h_{\hat{\mathbf{w}}}\right) + \tilde{O}\left(\sqrt{\frac{q}{m}}\right)\right) \geq 1 - \delta.$$

By the construction of $\tilde{\mathbf{w}}$ we know that $\hat{L}_0(h_{\tilde{\mathbf{w}}}) \leq \hat{L}_\gamma(h_{\mathbf{w}})$ which when combined with the above conclusion completes the proof of the theorem. □

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 machine learning given by [18]. This will provide some context to the framework under which PAC-Bayes bounds are derived. As before we suppose that our training data $S = \{(x_i, y_i)\}_{i=1}^m$ 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, $\mathcal{W}$. The learning algorithm requires some prior distribution, π, to be defined on $\mathcal{W}$. Then using the training data the posterior distribution, ρ, 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 realization of ρ.

3.1.2 Notations and Definitions

Bayesian machine learning is a way to manage randomness and uncertainty in the learning task. PAC-Bayes bounds are derived under this framework.

Definition 3.1 ([27])**.** *Let $\mathcal{M}(\mathcal{W})$ be a set of probability distributions defined over $\mathcal{W}$. A data-dependent probability measure is a function*

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

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

Definition 3.2 ([27])**.** *Given two probability measures Q and P defined on some sample space $\mathcal{X}$, the KL divergence between Q and P is*

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

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

Remark 3.3 ([15])**.** *When Q, P are probability measures on Euclidean space $\mathbb{R}^d$ with densities q, p respectively. The KL divergence is*

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

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

For the multivariate normal distributions [15] $N_q \sim \mathcal{N}(\mu_q, \Sigma_q)$ and $N_p \sim \mathcal{N}(\mu_p, \Sigma_p)$ defined on $\mathbb{R}^d$ we have that,

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

Similarly, for Bernoulli distributions [15] $\mathcal{B}(q) \sim \mathrm{Bern}(q)$ and $\mathcal{B}(p) \sim \mathrm{Bern}(p)$ it follows that

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

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

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

For a distribution Q defined on $\mathcal{W}$ we will use the notation

$$\mathbb{E}_{\mathbf{w}\sim Q}(R(\mathbf{w})) = R(Q) \text{ and } \mathbb{E}_{\mathbf{w}\sim Q}\left(\hat{R}(\mathbf{w})\right) = \hat{R}(Q)$$

for convenience. 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, $\pi \in \mathcal{M}(\mathcal{W})$.

3.1.3 PAC-Bayes Bounds

Theorem 3.4 ([8]). *For all $\lambda > 0$, for all $\rho \in \mathcal{M}(\mathcal{W})$, and $\delta \in (0,1)$ it follows that*

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

Theorem 3.4.1 (Jensen's Inequality). *For a convex function $f(x)$ (with a Taylor expansion) and a random variable X defined on sample space $\mathcal{X}$, if $\mathbb{E}(f(X))$ and $f(\mathbb{E}(X))$ are finite then*

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X)).$$

Equality holds if and only if f is a linear function on some convex set A such that $\mathbb{P}(X \in A) = 1$. If f doesn't have this property then equality holds if and only if the random variable is constant.

Proof ([5]). Let $\mu = \mathbb{E}(X)$, so by assumption we know there is a c such that

$$f(x) = f(\mu) + f'(\mu)(x-\mu) + \frac{f''(c)(x-\mu)^2}{2} \geq f(\mu) + f'(\mu)(x-\mu).$$

Where we have used the fact that $f''(c) > 0$ due to the convexity of f. Taking expectations of both sides we conclude that

$$\begin{aligned}\mathbb{E}(f(X)) &\geq \mathbb{E}\left(f(\mu) + f'(u)(X-\mu)\right)\\ &= \mathbb{E}(f(\mu)) + f'(\mu)\left(\mathbb{E}(X) - \mu\right)\\ &= f(\mu) = f\left(\mathbb{E}(X)\right),\end{aligned}$$

which completes the proof of the theorem. ■

Proposition 3.4.2. *For any probability measures Q and P it follows that $\mathrm{KL}(Q,P) \geq 0$ with equality if and only if Q and P are the same probability distribution.*

Proof. Note that $\log$ is a concave function so Jensen's inequality is reversed. Therefore,

$$\begin{aligned}-\mathrm{KL}(Q,P) &= -\int_{\mathcal{X}} \log\left(\frac{q(x)}{p(x)}\right) q(x)dx\\ &= \int_{\mathcal{X}} \log\left(\frac{p(x)}{q(x)}\right) q(x)dx\\ &= \mathbb{E}_Q\left(\log\left(\frac{p(x)}{q(x)}\right)\right)\\ &\leq \log\left(\mathbb{E}_Q\left(\frac{p(x)}{q(x)}\right)\right)\\ &= \log\left(\int_{\mathcal{X}} p(x)dx\right)\\ &= \log(1) = 0,\end{aligned}$$

where Jensen's inequality has been used to get the inequality. This shows that $\mathrm{KL}(Q,P) \geq 0$. Note that if $\mathrm{KL}(Q,P) = 0$ then equality must hold for Jensen's inequality which implies that $\frac{q(x)}{p(x)} = 1$ which implies that Q and P are the same probability distribution. On the other hand, if Q and P are the same probability distribution on the sample space $\mathcal{X}$ then,

$$\mathrm{KL}(Q,P) = \int_{\mathcal{X}} \log\left(\frac{q(x)}{p(x)}\right) q(x)dx = \int_{\mathcal{X}} \log(1)q(x)dx = 0.$$

■

Lemma 3.4.3. *For any measurable, bounded function $f : \mathcal{W} \to \mathbb{R}$ we have,*

$$\log\left(\mathbb{E}_{\mathbf{w}\sim\pi}\left(e^{f(\mathbf{w})}\right)\right) = \sup_{\rho\in\mathcal{M}(\mathcal{W})} \left(\mathbb{E}_{\mathbf{w}\sim\rho}\left(f(\mathbf{w})\right) - \mathrm{KL}(\rho,\pi)\right).$$

Moreover, the supremum with respect to ρ is achieved for the Gibbs posterior π_f defined by its density with respect to π as

$$\frac{d\pi_f(\mathbf{w})}{d\pi_f(\mathbf{w})} = \frac{e^{f(\mathbf{w})}}{\mathbb{E}_{\mathbf{w}\sim\pi_f}\left(e^{f(\mathbf{w})}\right)}.$$

Proof. From the definition of $\pi_f(\mathbf{w})$ we have that

$$\pi_f(\mathbf{w}) = \frac{e^{f(\mathbf{w})}}{\mathbb{E}_{\mathbf{w}\sim\pi_f}\left(e^{f(\mathbf{w})}\right)}\pi_f(\mathbf{w}).$$

Therefore,

$$\begin{aligned}
\mathrm{KL}\left(\rho,\pi_f\right) &= \int_{\mathbf{w}\in\mathcal{W}} \log\left(\frac{\rho(\mathbf{w})}{\pi_f(\mathbf{w})}\right)\rho(\mathbf{w})d\mathbf{w} \\
&= \int_{\mathbf{w}\in\mathcal{W}} \log(\rho(\mathbf{w}))\rho(\mathbf{w})d\mathbf{w} - \int_{\mathbf{w}\in\mathcal{W}} \log\left(\frac{e^{h(\mathbf{w})}\pi_f(\mathbf{w})}{\mathbb{E}_{\mathbf{w}\sim\pi_f}\left(e^{f(\mathbf{w})}\right)}\right)\rho(\mathbf{w})d\mathbf{w} \\
&= \int_{\mathbf{w}\in\mathcal{W}} \log\left(\frac{\rho(\mathbf{w})}{\pi_f(\mathbf{w})}\right)\rho(\mathbf{w})d\mathbf{w} - \int_{\mathbf{w}\in\mathcal{W}} h(\mathbf{w})\rho(\mathbf{w})d\mathbf{w} + \log\left(\mathbb{E}_{\mathbf{w}\sim\pi_f}\left(e^{f(\mathbf{w})}\right)\right) \\
&= \mathrm{KL}(\rho,\pi_f) - \mathbb{E}_\rho(f(\mathbf{w})) + \log\left(\mathbb{E}_{\mathbf{w}\sim\pi_f}\left(e^{f(\mathbf{w})}\right)\right).
\end{aligned}$$

By Proposition 3.4.2 the left hand side is non-negative and equal to 0 only when $\rho = \pi_f$, which completes the proof. ■

Proof. Recall, from the proof of Theorem 2.1 that for any $t > 0$ we have that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(tm\left(R(\mathbf{w}) - \hat{R}(\mathbf{w})\right)\right)\right) \leq \exp\left(\frac{mt^2C^2}{8}\right).$$

Letting $t = \frac{\lambda}{m}$ we deduce that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(R(\mathbf{w}) - \hat{R}(\mathbf{w})\right)\right)\right) \leq \exp\left(\frac{\lambda^2C^2}{8m}\right).$$

Integrating this with respect to π gives

$$\mathbb{E}_{\mathbf{w}\sim\pi}\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(R(\mathbf{w}) - \hat{R}(\mathbf{w})\right)\right)\right) \leq \exp\left(\frac{\lambda^2C^2}{8m}\right).$$

To which we can apply Fubini's theorem to interchange the order of integration

$$\mathbb{E}_{S\sim\mathcal{D}^m}\exp\left(\lambda\left(R(\pi)-\hat{R}(\pi)\right)\right)\leq\exp\left(\frac{\lambda^2C^2}{8m}\right),$$

and then apply Lemma 3.4.3 to get

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\sup_{\rho\in\mathcal{M}(\mathcal{W})}\left(\lambda\left(R(\rho)-\hat{R}(\rho)\right)\right)-\mathrm{KL}(\rho,\pi)-\frac{\lambda^2C^2}{8m}\right)\right)\leq 1.$$

Now fix $s>0$ and apply Chernoff bound to get that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\sup_{\rho\in\mathcal{M}(\mathcal{W})}\left(\lambda\Big(R(\rho)-\hat{R}(\rho)\Big)\right)-\mathrm{KL}(\rho,\pi)-\frac{\lambda^2C^2}{8m}>s\right)$$
$$\leq\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\sup_{\rho\in\mathcal{M}(\mathcal{W})}\left(\lambda\left(R(\rho)-\hat{R}(\rho)\right)\right)-\mathrm{KL}(\rho,\pi)\right)\right)e^{-s}\leq e^{-s}.$$

Setting $s=\log\left(\frac{1}{\delta}\right)$ and rearranging completes the proof. □

Theorem 3.4 motivates the study of the data-dependent probability measure

$$\hat{\rho}_\lambda=\mathrm{argmin}_{\rho\in\mathcal{M}(\mathcal{W})}\left(\hat{R}(\rho)+\frac{\mathrm{KL}(\rho,\pi)}{\lambda}\right). \tag{1}$$

Definition 3.5 ([27]). *The optimization problem defined by Equation (1) has the solution $\hat{\rho}_\lambda=\pi_{-\lambda\hat{R}}$ given by*

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

This is distribution is known as the Gibbs posterior.

Corollary 3.6 ([27]). *For all $\lambda>0$, and $\delta\in(0,1)$ it follows that*

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

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 realization of the posterior distribution, we have the following result.

Theorem 3.7. *[27] For all $\lambda>0$, $\delta\in(0,1)$, and data-dependent probability measure $\tilde{\rho}$ we have that*

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

Proof. The beginning of this proof proceeds in the same way as that of Theorem 3.4 up to the point where we conclude that

$$\mathbb{E}_{\mathbf{w}\sim\pi}\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(R(\mathbf{w})-\hat{R}(\mathbf{w})\right)\right)\right)\leq\exp\left(\frac{\lambda^2C^2}{8m}\right).$$

For any non-negative function h we have that

$$\begin{aligned}\mathbb{E}_{\mathbf{w}\sim\pi}(h(\mathbf{w})) &= \int_{\mathbf{w}\in\mathcal{W}} h(\mathbf{w})\pi(d\mathbf{w}) \\ &= \int_{\left\{\frac{d\tilde{\rho}}{d\pi}(\mathbf{w})>0\right\}} h(\mathbf{w})\pi(d\mathbf{w}) \\ &= \int_{\left\{\frac{d\tilde{\rho}}{d\pi}(\mathbf{w})>0\right\}} h(\mathbf{w})\frac{d\pi}{d\tilde{\rho}}(\mathbf{w})\tilde{\rho}(d\mathbf{w}) \\ &= \mathbb{E}_{\mathbf{w}\sim\tilde{\rho}}\left(h(\mathbf{w})\exp\left(-\log\left(\frac{d\tilde{\rho}}{d\pi}(\mathbf{w})\right)\right)\right)\end{aligned}$$

which means that

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

Now in the same way as the proof of Theorem 3.4 we apply the Chernoff bound, set δ and then re-arrange the terms to complete the proof. □

Note that Theorem 3.4 is a bound in probability. We now state an equivalent bound that holds in expectation.

Theorem 3.8. *[27] For all $\lambda > 0$, and data-dependent probability measure $\tilde{\rho}$, we have that*

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

Proof. Once again we proceed in the same way as Theorem 3.4 to the point where we deduce that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\sup_{\rho\in\mathcal{M}(\mathcal{W})}\left(\lambda\left(R(\rho)-\hat{R}(\rho)\right)\right)-\mathrm{KL}(\rho,\pi)-\frac{\lambda^2C^2}{8m}\right)\right)\leq 1.$$

Now we apply Jensen's inequality to get that

$$\exp\left(\mathbb{E}_{S\sim\mathcal{D}^m}\left(\sup_{\rho\in\mathcal{M}(\mathcal{W})}\left(\lambda\left(R(\rho)-\hat{R}(\rho)\right)\right)-\mathrm{KL}(\rho,\pi)-\frac{\lambda^2C^2}{8m}\right)\right)\leq 1,$$

which implies that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\sup_{\rho\in\mathcal{M}(\mathcal{W})}\left(\lambda\left(R(\rho)-\hat{R}(\rho)\right)\right)-\mathrm{KL}(\rho,\pi)-\frac{\lambda^2C^2}{8m}\right)\leq 0.$$

In particular this holds for our data-dependent probability measure $\tilde{\rho}$. Therefore,

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\lambda\left(R(\tilde{\rho})-\hat{R}(\tilde{\rho})\right)-\mathrm{KL}(\tilde{\rho},\pi)-\frac{\lambda^2C^2}{8m}\right)\leq 0,$$

and so using the linearity of expectation and rearranging completes the proof. □

Corollary 3.9 ([27])**.** *For $\tilde{\rho}=\hat{\rho}_\lambda$, the following holds*

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

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

Theorem 3.10. *[2] For all $\rho \in \mathcal{M}(\mathcal{W})$ and $\delta > 0$ we have that*

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

Proof. Refer to (McAllester, 1999) for the proof of this theorem. □

Theorem 3.11 ([7])**.** *For $a > 0$ and $p \in (0, 1)$ let*

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

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

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

Proof. Refer to (Catoni, 2007) for the proof of this theorem. □

Theorem 3.12 ([4])**.** *For any $\delta > 0$ and $\rho \in \mathcal{M}(\mathcal{W})$ then we have that*

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

For $X_1, \ldots, X_n$ i.i.d random variables in $[0, 1]$ and with $\mathbb{E}(X_i) = \mu$ let $\mathbf{X} = (X_1, \ldots, X_n)$ and

$$M(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

For any random variable X in $[0, 1]$ let X' denote the Bernoulli random variables with parameter $\mathbb{E}(X)$ and let $\mathbf{X}' = (X'_1, \ldots, X'_n)$.

Theorem 3.12.1. *For $n \geq 2$ with the notation as above we have that*

$$\mathbb{E}\left(\exp\left(n\mathrm{kl}(M(\mathbf{X}), \mu)\right)\right) \leq \exp\left(\frac{1}{12n}\right)\sqrt{\frac{\pi n}{2}} + 2.$$

Proof. For the proof of this theorem refer to (Maurer, 2004). ■

Corollary 3.12.2. *For $n \geq 2$ we have that*

$$\mathbb{E}\left(\exp\left(n\mathrm{kl}(M(\mathbf{X}), \mu)\right)\right) \leq 2\sqrt{n}.$$

Proof. Replace n with the continuous variable $x \in (0, \infty)$. Let $f(x) = \exp\left(\frac{1}{12x}\right)\sqrt{\frac{\pi x}{2}} + 2$ and $g(x) = 2\sqrt{x}$, then

$$f'(x) = g'(x)\left(\sqrt{\frac{\pi}{2}}\exp\left(\frac{1}{12x}\right)\left(\frac{1}{2} - \frac{1}{12x}\right)\right).$$

From which it is clear that $f'(x) < g'(x)$. Therefore, as one can numerically see that $g(x) > f(x)$ for $x \approx 7.5$ we can conclude that for all $n \geq 8$ we have that $\exp\left(\frac{1}{12n}\right)\sqrt{\frac{\pi n}{2}} + 2 \leq 2\sqrt{n}$ which completes

the proof of the corollary. ■

Proof. Recall, that

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

and $R(\mathbf{w}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left(l(h(x), y)\right)$. As we are considering a loss function bounded to the interval $[0, 1]$ we can consider each of the $l(h_{\mathbf{w}}(x_i), y_i)$ as i.i.d random variables with mean $R(\mathbf{w})$. Therefore, for any $\mathbf{w} \in \mathcal{W}$ we can apply Corollary 3.12.2 to deduce that

$$\mathbb{E}\left(m\mathrm{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right) \leq 2\sqrt{m}.$$

Now applying Jensen's inequality to the convexity of kl divergence and the exponential function we have that

$$\begin{aligned}
\mathbb{E} - S \sim \mathcal{D}^m\left(\exp\left(m\mathrm{kl}\left(\hat{R}(\rho), R(\rho)\right) - \mathrm{kl}\left(\rho, \pi\right)\right)\right) &\leq \mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\mathbb{E}_{\mathbf{w}\sim\rho}\left(m\mathrm{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right) - \log\left(\frac{d\rho(\mathbf{w})}{d\pi(\mathbf{w})}\right)\right)\right)\right) \\
&\leq \mathbb{E}_{S\sim\mathcal{D}^m}\left(\mathbb{E}_{\mathbf{w}\sim\rho}\left(\exp\left(m\mathrm{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right) - \log\left(\frac{d\rho(\mathbf{w})}{d\pi(\mathbf{w})}\right)\right)\right)\right) \\
&= \mathbb{E}_{S\sim\mathcal{D}^m}\left(\mathbb{E}_{\mathbf{w}\sim\pi}\left(\exp\left(m\mathrm{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right)\left(\frac{d\rho}{d\pi}\right)^{-1}\frac{d\rho}{d\pi}\right)\right) \\
&\leq \mathbb{E}_{\mathbf{w}\sim\rho}\left(\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(m\mathrm{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right)\right)\right) \\
&\leq 2\sqrt{m}.
\end{aligned}$$

$$Applying Markov's inequality we conclude that$$

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}\left(\mathrm{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right) > \frac{\mathrm{kl}(\rho,\pi) + \log\left(\frac{2\sqrt{m}}{\delta}\right)}{m}\right) &= \mathbb{P}_{S\sim\mathcal{D}^m}\left(\exp\left(m\mathrm{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right) - \mathrm{kl}(\rho,\pi)\right) > \frac{2\sqrt{m}}{\delta}\right) \\
&\leq \delta.
\end{aligned}$$

Taking the complement of this completes the proof. □

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 [15] 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 $\hat{R}(\mathbf{w}) \approx \hat{R}(\mathbf{w}^*)$ [27]. The setup considered by [15] is the same as the one we have considered throughout this report. With $\mathcal{X} \subset \mathbb{R}^k$ and labels being ± 1. That is, we are considering binary classification based on a set of features. We explicitly state our hypothesis set as

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

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 $l : \mathbb{R} \to \{\pm 1\}$ be defined as $l(y, y') = \mathbf{1}(\mathrm{sgn}(y') = y)$. For

optimization purposes we use the convex surrogate loss function $\tilde{l} : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}_+$

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

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. As noted previously the work [15] 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 minimize the divergence between the prior and the posterior when the learning algorithm is stochastic gradient descent (SGD). To begin [15] utilize the following bound.

Theorem 3.13 ([15])**.** *For every $\delta > 0$,$m \in \mathbb{N}$, distribution $\mathcal{D}$ on $\mathbb{R}^k \times \{\pm 1\}$, distribution π on $\mathcal{W}$ and distribution $\rho \in \mathcal{M}(\mathcal{W})$, we have that*

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

Remark 3.14. *Note how this is a slightly weaker statement than Theorem 3.12. This is because [15] cited this Theorem from [3], however, since then [4] was able to tighten the result by providing Theorem 3.12. In the following we will update the work of [15] and use the tightened result provided by Theorem 3.12.*

This motivates the following PAC-Bayes learning algorithm.

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

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

- Then return the randomized classifier given by ρ.

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 biases of our neural network. To make the algorithm more practical we therefore consider

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

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

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

We now suppose our prior π 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 λ we discretize the problem, with the side-effect of expanding the eventual generalization bound. We let λ have the for $c\exp\left(-\frac{j}{b}\right)$ for $j \in \mathbb{N}$, so that c is an upper bound and b controls precision. By ensuring that Theorem 3.12 holds with probability

$1 - \frac{6\delta}{\pi^2 j^2}$ for each $j \in \mathbb{N}$ we can then apply a union bound argument to ensure that we get results that hold for probability $1 - \delta$. Treating λ as continuous during the optimization process and then discretized at the point of evaluating the bound yields the updated optimization problem

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

where

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

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 $\tilde{R}\left(\mathbf{w} + \xi \odot \sqrt{\mathbf{s}}\right)$ where $\xi \sim \mathcal{N}_{0,\mathbf{1}_d}$. Once good candidates for this optimization problem are found we return to (2) to calculate the final error bound. With the choice of λ it follows that with probability $1-\delta$, uniformly over all $\mathbf{w} \in \mathbb{R}^d, \mathbf{s} \in \mathbb{R}^d_+$ and λ (of the discrete form) the expected risk of $\rho = \mathcal{N}_{\mathbf{w},\mathbf{s}}$ is bounded by

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

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

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

which holds with probability $1 - \delta'$. Finally, by the union bound

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

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 λ that give a tighter bound.

Once the values of $\mathbf{w}, \mathbf{s}$ and λ are found we then need to compute $\overline{\hat{R}_{n,\delta'}}(\rho) := \mathrm{kl}^{-1}\left(\hat{R}\left(\hat{\rho}_n\right), \frac{1}{n}\log\left(\frac{2}{\delta'}\right)\right)$ 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 λ.
$\tau \in (0,1), T$, learning rate.
Ensure: Optimal $\mathbf{w}, \mathbf{s}, \lambda$.

$\zeta = |\mathbf{w}|$ ▷ $\mathbf{s}(\zeta) = e^{2\zeta}$
$\rho = -3$ ▷ $\lambda(\rho) = e^{2\rho}$
$B(\mathbf{w}, \mathbf{s}, \lambda, \mathbf{w}') = \tilde{R}(\mathbf{w}) + \sqrt{\frac{1}{2} B_{\mathrm{RE}}(\mathbf{w}, \mathbf{s}, \lambda)}$
for $t = 1 \to T$ **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 $p \approx \mathrm{kl}^{-1}(q, c)$
for $n = 1 \to N$ **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 behavior of the bounds and they can take the form

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

Where $r_m(\delta)$ is a remainder term that tends to 0 as m tends to ∞. Although this bound cannot be computed in practice it is illustrative of the behavior 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 ([27])**.** *For $\lambda > 0$ we have that*

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

Theorem 4.1.1 (Fubini's Theorem)**.** *If $\mathcal{X}_1$ and $\mathcal{X}_2$ are σ-finite measurable spaces and $f : \mathcal{X}_1 \times \mathcal{X}_2$ is measurable and*

$$\int_{\mathcal{X}_1\times\mathcal{X}_2} |f(x_1,x_2)| d(x_1,x_2) < \infty,$$

then

$$\int_{\mathcal{X}_1}\left(\int_{\mathcal{X}_2} f(x_1,x_2)dx_2\right)dx_1 = \int_{\mathcal{X}_2}\left(\int_{\mathcal{X}_1} f(x_1,x_2)dx_2\right)dx_1 = \int_{\mathcal{X}_1\times\mathcal{X}_2} f(x_1,x_2)d(x_1,x_2).$$

Proof. For the proof of this theorem please refer to [23]. ■

Proof. We proceed from Corollary 3.9 to deduce that

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

where Fubini's theorem has been applied in the last inequality. Recalling that $\mathbb{E}_{S\sim\mathcal{D}^m}\left(\hat{R}(\mathbf{w})\right) = R(\mathbf{w})$ completes the proof of the theorem. □

4.1.2 Oracle PAC-Bayes Bounds in Probability

Theorem 4.2 ([27])**.** *For any $\lambda > 0$, and $\delta \in (0,1)$ we have that*

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

Proof. Recall the proof of Theorem 3.4 and the subsequent application to the Gibbs posterior that yielded Corollary 3.6.

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

In the proof we utilized the result of Theorem 2.1. The inequality of Theorem 2.1 can be reversed by replacing the U_i by $-U_i$ in its proof. Applying the reverse inequality of Theorem 2.1 in the proof of Theorem 3.4 gives the updated corollary

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

Which holds for all $\rho \in \mathcal{M}(\mathcal{W})$. Applying a union bound on Corollary 3.6 and the updated result above gives

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\begin{array}{c} R(\hat{\rho}_\lambda) \leq \inf_{\rho\in\mathcal{M}(\mathcal{W})}\left(\hat{R}(\rho) + \frac{\lambda C^2}{8m} + \frac{\mathrm{KL}(\rho,\pi)+\log\left(\frac{1}{\delta}\right)}{\lambda}\right), \\ \hat{R}(\rho) \leq R(\rho) + \frac{\lambda C^2}{8m} + \frac{\mathrm{KL}(\rho,\pi)+\log\left(\frac{1}{\delta}\right)}{\lambda} \end{array}\right) \geq 1-2\delta,$$

which holds for all $\rho \in \mathcal{M}(\mathcal{W})$. Using the upper bound on $\hat{R}(\rho)$ from the second event on the first event gives

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

We can simply replace the δ with $\frac{\delta}{2}$ to complete the proof. □

4.1.3 Bernstein's Assumption

Definition 4.3 ([27]). *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

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

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

Theorem 4.4 ([27]). *Assume Bernstein's assumption is satisfied with some constant $K > 0$. Take $\lambda = \frac{m}{\max(2K,C)}$ then we have*

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

Lemma 4.4.1. *Let g denote the Bernstein function defined by*

$$g(x) = \begin{cases} 1 & x = 0 \\ \frac{e^x - 1 - x}{x^2} & x \neq 0. \end{cases}$$

Let $U_1, \ldots, U_n$ be i.i.d *random variables such that $\mathbb{E}(U_i)$ is finite and $U_i - \mathbb{E}(U_i) \leq C$ almost surely for some $C \in \mathbb{R}$. Then,*

$$\mathbb{E}\left(\exp\left(t\sum_{i=1}^{n}\left(U_i - \mathbb{E}(U_i)\right)\right)\right) \leq \exp\left(g(Ct)nt^2\mathrm{Var}(U_i)\right).$$

Proof ([1]). We first show that function g is increasing. For $x \neq 0$ we have that

$$g'(x) = \frac{(x-2)e^x + 2 + x}{x^3}.$$

Let $h(x) = (x-2)e^x + 2 + x$ then $h(0) = 0$ and $h'(x) = (x-2)e^x + 1$, so that $h'(0) = 0$ and $h''(x) = xe^x$. Therefore, $h'(x) < 0$ for $x < 0$ and $h'(x) > 0$ for $x > 0$ which implies that $h(x) \geq 0$ for all x. This means that $g'(x) > 0$ and the function g is increasing. So that

$$e^x = 1 + x + x^2 g(x) \leq 1 + x + x^2 g(\alpha)$$

for $x \leq \alpha$. Therefore, if we have a random variable X with $\mathbb{E}(X) = 0$ and $X \leq \alpha$ it follows that

$$\mathbb{E}\left(\exp(X)\right) \leq 1 + g(\alpha)\mathrm{Var}(X) \leq \exp(g(\alpha)\mathrm{Var}(X)).$$

Applying this conclusion to $\alpha = Ct$, $X = t(U_i - \mathbb{E}(U_i))$ we can conclude that

$$\mathbb{E}\left(\exp\left(t(U_i - \mathbb{E}(U_i))\right)\right) \leq \exp\left(g(Ct)t^2\mathrm{Var}(U_i)\right)$$

Therefore, by the independence of the U_i

$$\begin{aligned}
\mathbb{E}\left(\exp\left(t\sum_{i=1}^{n}\left(U_i - \mathbb{E}(U_i)\right)\right)\right) &= \prod_{i=1}^{n}\mathbb{E}\left(\exp\left(t(U_i - \mathbb{E}(U_i))\right)\right) \\
&\leq \prod_{i=1}^{n}\exp\left(g(Ct)t^2\mathrm{Var}(U_i)\right) \\
&= \exp\left(g(Ct)nt^2\mathrm{Var}(U_i)\right)
\end{aligned}$$

as required. ■

Proof. Now fix $\mathbf{w} \in \mathcal{W}$ and apply Lemma 4.4.1 to $U_i = l_i(\mathbf{w}^*) - l_i(\mathbf{w})$ (where we inherit the notation of the proof of Theorem 2.1). Note that $\mathbb{E}(U_i) = R(\mathbf{w}^*) - R(\mathbf{w})$ and therefore,

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(tm\left(R(\mathbf{w}) - R(\mathbf{w}^*) - \hat{R}(\mathbf{w}) + \hat{R}(\mathbf{w})\right)\right)\right) \leq \exp\left(g(Ct)mt^2\mathrm{Var}_{S\sim\mathcal{D}^m}(U_i)\right).$$

Observe that

$$\begin{aligned}
\mathrm{Var}(U_i) &\leq \mathbb{E}_{S\sim\mathcal{D}^m}\left(U_i^2\right) \\
&= \mathbb{E}_{S\sim\mathcal{D}^m}\left(l_i(\mathbf{w}^*) - l_i(\mathbf{w})\right) \\
&\leq K(R(\mathbf{w}) - R(\mathbf{w}^*)).
\end{aligned}$$

Therefore, with $\lambda = tn$ we get that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(R(\mathbf{w}) - R(\mathbf{w}^*) - \hat{R}(\mathbf{w}) + \hat{R}(\mathbf{w}^*)\right)\right)\right) \leq \exp\left(g\left(\frac{\lambda C}{m}\right)\frac{\lambda^2}{m}K\left(R(\mathbf{w}) - R(\mathbf{w}^*)\right)\right)$$

which upon rearrangement gives

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(1 - Kg\left(\frac{\lambda C}{m}\right)\frac{\lambda}{m}\right)\left(R(\mathbf{w}) - R(\mathbf{w}^*)\right) - \hat{R}(\mathbf{w}) - \hat{R}(\mathbf{w}^*)\right)\right) \leq 1.$$

Now integrate with respect to π and apply Fubini's theorem along with Lemma 3.4.3 from the proof of

Theorem 3.4 to get

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\sup_{\rho\in\mathcal{M}(\mathcal{W})}\left(\left(1-Kg\left(\frac{\lambda C}{m}\right)\frac{\lambda}{m}\right)(R(\rho)-R(\mathbf{w}^*))-\hat{R}(\rho)-\hat{R}(\mathbf{w}^*)-\mathrm{KL}(\rho,\pi)\right)\right)\right)\leq 1.$$

In particular, this holds for $\rho-\hat{\rho}_\lambda$, and we can apply Jensen's inequality and re-arrange to yield

$$\left(1-Kg\left(\frac{\lambda C}{m}\right)\right)\left(\mathbb{E}_{S\sim\mathcal{D}^m}\left(R(\hat{\rho}_\lambda)-R(\mathbf{w}^*)\right)\right)\leq\mathbb{E}_{S\sim\mathcal{D}^m}\left(\hat{R}(\rho)-\hat{R}(\mathbf{w})+\frac{\mathrm{KL}(\hat{\rho}_\lambda,\pi)}{\lambda}\right).$$

From now on λ will be such that $1-Kg\left(\frac{\lambda C}{m}\right)\frac{\lambda}{m}>0$, thus

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(R(\hat{\rho}_\lambda)\right)-R(\mathbf{w}^*)\leq\frac{\mathbb{E}_{S\sim\mathcal{D}^m}\left(\hat{R}(\hat{\rho}_\lambda)-\hat{R}(\mathbf{w}^*)+\frac{\mathrm{KL}(\hat{\rho}_\lambda,\pi)}{\lambda}\right)}{1-Kg\left(\frac{\lambda C}{m}\right)\frac{\lambda}{m}}.$$

As with $\lambda=\frac{m}{\max(2K,C)}$ it follows that

$$Kg\left(\frac{\lambda C}{m}\right)\frac{\lambda}{m}\leq\frac{1}{2}$$

and so we have

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

As $\hat{\rho}_\lambda$ minimizes the quantity on the right hand side in expectation we can re-write this as

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

which completes the proof. □

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 [20] 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 ([10])**.** *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*

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

Lemma 4.5.1. *For $\lambda>\frac{1}{2}$, if $\mathrm{kl}_{-\frac{1}{\gamma}}(q,p)\leq c$ then*

$$p\leq\frac{1}{1-\frac{1}{2\lambda}}(q+\lambda c).$$

Proof. Let $\gamma = -\frac{1}{\lambda}$ for convenience, which means that $\gamma \in (-2, 0)$. Re-arranging the assumption we get that

$$p \leq \frac{1 - e^{\gamma q - c}}{1 - e^{\gamma}}.$$

Using $e^{\gamma} \geq 1 + \gamma$ in the numerator and $e^{\gamma} \leq 1$ in the denominator we get

$$p \leq \frac{q - \frac{c}{\gamma}}{1 + \frac{1}{2}\gamma},$$

which when we substitute λ back in completes the proof of the lemma. ■

Lemma 4.5.2. *Let $x_1, \ldots, x_n$ be realizations of a random variable X with range $[0, 1]$ and mean μ. Let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then for any fixed γ we have that*

$$\mathbb{E}\left(\exp\left(n\mathrm{kl}_{\gamma}(\hat{\mu}, \mu)\right)\right) \leq 1.$$

Proof. Note that $\mathbb{E}\left(\exp(n\gamma\hat{\mu})\right) = \left(\mathbb{E}(\exp(\gamma X))\right)^n$ and that by the convexity of $\exp(\cdot)$ we have that

$$e^{\gamma X} \leq 1 - x + xe^{\gamma}.$$

Therefore,

$$\mathbb{E}\left(\exp\left(n\gamma\hat{\mu}\right)\right) \leq \left(1 - \mu + \mu e^{\gamma}\right)^n,$$

which implies that

$$\mathbb{E}\left(\exp\left(n\left(\gamma\hat{\mu} - \log\left(1 - \mu + \mu e^{\gamma}\right)\right)\right)\right) \leq 1$$

which completes the proof of the lemma. ■

Lemma 4.5.3. *For probability distributions defined on the sample space $\mathcal{X}$ and a measurable function f we have that*

$$\mathbb{E}_{x\in Q}(f(x)) \leq \mathrm{KL}(Q, P) + \log\left(\mathbb{E}_{x\in P}\left(\exp(f(x))\right)\right).$$

Proof.

$$\begin{aligned}
\mathbb{E}_{x\in Q}\left(f(x)\right) &= \mathbb{E}_{x\in Q}\left(\log\left(\exp(f(x))\right)\right) \\
&= \mathbb{E}_{x\in Q}\left(\log\left(\frac{P(x)}{Q(x)}\right)e^{f(x)} + \log\left(\frac{Q(x)}{P(x)}\right)\right) \\
&\leq \log\left(\mathbb{E}_{x\in Q}\left(\frac{P(x)}{Q(x)}e^{f(x)}\right)\right) + \mathrm{KL}(Q, P) \\
&= \mathrm{KL}(Q, P) + \log\left(\mathbb{E}_{x\in P}\left(\exp(f(x))\right)\right).
\end{aligned}$$

■

Proof. We can use similar reasoning to that given in the proof of Theorem 3.12 to conclude from Lemma 4.5.2 that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(m\mathrm{kl}_{\gamma}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right)\right) \leq 1$$

for fixed $\mathbf{w} \in \mathcal{W}$. Now we can take expectations over π on both sides an apply Fubini's theorem to

deduce that

$$1 \geq \mathbb{E}_{\mathbf{w}\sim\pi}\left(\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(m\mathrm{kl}_\gamma\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right)\right)\right)$$
$$\geq \mathbb{E}_{S\sim\mathcal{D}^m}\left(\mathbb{E}_{\mathbf{w}\sim\pi}\left(\exp\left(m\mathrm{kl}_\gamma\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right)\right)\right).$$

To which we can apply Markov's inequality to get that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\mathbb{E}_{\mathbf{w}\sim\pi}\left(\exp\left(m\mathrm{kl}_\gamma\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right)\right) \leq \frac{1}{\delta}\right) \geq 1-\delta.$$

Letting $f(\mathbf{w}) = m\mathrm{kl}_\gamma\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)$ in Lemma 4.5.3 and using the above result we get that

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

By the convexity of kl_γ we get that

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

Therefore, by re-arranging and applying Lemma 4.5.1 the proof of the theorem is complete. □

Corollary 4.6 ([20]). *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*

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

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

Proof. This is the result of the previous Theorem 4.5 with $\lambda = \frac{1}{2(1-\beta)}$ and $C = 1$. □

As we have done previously, we can consider the optimization problem of minimizing the bound of Corollary 4.6.

Theorem 4.7 ([20]). *Let $m \in \mathbb{N}$ and fix a probability kernel $\rho : \mathcal{Z}^m \to \mathcal{M}(\mathcal{W})$. Then for all $\beta, \delta \in (0,1)$ and distributions $\mathcal{D}$ defined on $\mathcal{Z}$ we that $\mathbb{E}_{S\sim\mathcal{D}^m}\left(\Psi_{\beta,\delta}(\rho(S),\pi;S)\right)$ is minimized, in π, by the oracle prior $\pi^* = \mathbb{E}_{S\sim\mathcal{D}^m}(\rho(S))$.*

For a subset J of $\{1,\ldots,m\}$ 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

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

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

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

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_α 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_α and U. Our aim now is to tighten the bound of Corollary 4.6 by minimizing $\mathbb{E}^{S_\alpha, U}(\mathrm{KL}(\rho(U, S), \pi))$. 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 σ fixed we are left with the minimization problem

$$\mathrm{argmin}_{\mathbf{w}_\alpha} \left(\mathbb{E}^{S_\alpha, U} \left(\|\mathbf{w}_S - \mathbf{w}_\alpha\| \right) \right),$$

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_α to obtain the sample S^G_α. Let $\mathbf{w}^G_\alpha$ be the mean of $\rho(U, S^G_\alpha)$. By construction, SGD will first process S_α then the combined portion of S^G and hence $\mathbf{w}^G_\alpha$ and $\mathbf{w}_S$ are equal in distribution when conditioned on S_α and U. Therefore, $\mathbf{w}^G_\alpha$ 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_α is the α-prefix run.
- The SGD run on S^G_α is the α-prefix+ghost run and obtains the parameters $\mathbf{w}^G_\alpha$.

The resulting parameters of the α-prefix and α-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 α-prefix runs to obtain the mean of the prior. It is not clear whether α-prefix+ghost run will always obtain a parameter that leads to a tighter generalization bound. Recall, that σ 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 η
  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 $l_S^{0\text{-}1}(\mathbf{w}) \leq \mathcal{E}$ **then**
        break
      end if
    end for
  end function

Algorithm 7 Obtaining Bound Using SGD Informed Prior

Require: Stopping criteria $\mathcal{E}$, Prefix fraction α, Ghost Data S^G (If available), Batch size b.

function GETBOUND($\mathcal{E}, \alpha, T, \sigma_P$)

$S_\alpha \leftarrow \{z_1, \ldots, z_{\alpha|S|} \subset S\}$

$\mathbf{w}_\alpha^0 \leftarrow$SGD$\left(\mathbf{w}_0, S_\alpha, b, \frac{|S_\alpha|}{b}\right)$

$\mathbf{w}_S \leftarrow$SGD$\left(\mathbf{w}_\alpha^0, S, b, \infty, \mathcal{E}\right)$ ▷ Base Run

$\mathbf{w}_\alpha^G \leftarrow$SGD$\left(\mathbf{w}_\alpha^0, S_\alpha^G, b, T, \cdot\right)$ ▷ Ghost run if data available, otherwise prefix run

$\pi \leftarrow \mathcal{N}\left(\mathbf{w}_\alpha^G, \sigma I\right)$

$\rho \leftarrow \mathcal{N}\left(\mathbf{w}_S, \sigma I\right)$

Bound$\leftarrow \Psi_\delta^*(\rho, \pi; S \setminus S_\alpha)$

return Bound

end function

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 [24] 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 $A : \mathcal{Z}^m \to \mathcal{M}(\mathcal{W})$ that is applied to the training sample S.

Definition 5.1 ([24]). *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 ([24]). *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any parameter space $\mathcal{W}$, for any prior distribution π on $\mathcal{W}$, for any $\phi : \mathcal{W} \times \mathcal{Z}^m \to \mathbb{R}^+$, for any $\alpha > 1$, for any $\delta > 0$ and for any deterministic learning algorithm $A : \mathcal{Z}^m \to \mathcal{M}(\mathcal{W})$ the following holds*

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

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

Proof. First note that $\phi(\mathbf{w}, S)$ is a non-negative random variable. Therefore, by Markov's inequality

$$\mathbb{P}_{\mathbf{w} \sim \rho_S} \left(\phi(\mathbf{w}, S) \leq \frac{2}{\delta} \mathbb{E}_{\mathbf{w}' \sim \rho_S} \left(\phi\left(\mathbf{w}', S \right) \right) \right) \geq 1 - \frac{\delta}{2},$$

which is equivalent to

$$\mathbb{E}_{\mathbf{w} \sim \rho_S} \left(\phi(\mathbf{w}, S) \leq \frac{2}{\delta} \mathbb{E}_{\mathbf{w}' \sim \rho_S} \left(\phi\left(\mathbf{w}', S \right) \right) \right) \geq 1 - \frac{\delta}{2}.$$

Taking the expectations over $S \sim \mathcal{D}^m$ to both we obtain the equivalent statements

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

and

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

Taking the $\log$ of the first of these and then multiplying by $\frac{\alpha}{\alpha-1}$ gives

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

Focusing on the right hand side we see that

$$\frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\mathbb{E}_{\mathbf{w}'\sim\rho_S}\left(\phi\left(\mathbf{w}',S\right)\right)\right)=\frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\mathbb{E}_{\mathbf{w}'\sim\rho_S}\left(\frac{\rho_S(\mathbf{w}')\pi(\mathbf{w}')}{\pi(\mathbf{w}')\rho_S(\mathbf{w}')}\phi(\mathbf{w}',S)\right)\right)$$

for all $\pi\in\mathcal{M}(\mathcal{W})$. As $\frac{1}{\alpha}+\frac{1}{\frac{\alpha}{\alpha-1}}=1$ we can apply Holder's inequality to get that

$$\mathbb{E}_{\mathbf{w}'\sim\pi}\left(\frac{\rho_S(\mathbf{w}')}{\pi(\mathbf{w}')}\phi\left(\mathbf{w}',S\right)\right)\leq\left(\mathbb{E}_{\mathbf{w}'\sim\pi}\left(\frac{\rho_S(\mathbf{w}')}{\pi(\mathbf{w}')}\right)^{\alpha}\right)^{\frac{1}{\alpha}}\left(\mathbb{E}_{\mathbf{w}'\sim\pi}\left(\phi(\mathbf{w}',S)^{\frac{\alpha}{\alpha-1}}\right)\right)^{\frac{\alpha-1}{a}}.$$

Therefore,

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

From which we deduce that

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

As $\mathbb{E}_{\mathbf{w}'\sim\pi}\phi(\mathbf{w}',S)^{\frac{\alpha}{\alpha-1}}$ is also a non-negative random variables we can apply Markov's inequality again to get

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

As the left hand side is not dependent of $\mathbf{w}\sim\rho_S$ we have that

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}\Bigg(&\mathbb{E}_{\mathbf{w}'\sim\pi}\phi(\mathbf{w}',S)^{\frac{\alpha}{\alpha-1}}\leq\frac{\delta}{2}\mathbb{E}_{S'\mathcal{D}^m}\mathbb{E}_{\mathbf{w}'\sim\pi}\phi(\mathbf{w}',S')^{\frac{\alpha}{\alpha-1}}\Bigg)\\&=\mathbb{P}_{S\sim\mathcal{D}^m,\mathbf{w}\sim\rho_S}\left(\mathbb{E}_{\mathbf{w}'\sim\pi}\phi(\mathbf{w}',S)^{\frac{\alpha}{\alpha-1}}\leq\frac{\delta}{2}\mathbb{E}_{S'\mathcal{D}^m}\mathbb{E}_{\mathbf{w}'\sim\pi}\phi(\mathbf{w}',S')^{\frac{\alpha}{\alpha-1}}\right).\end{aligned}$$

Therefore,

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

Combining with $(\star)$ using a union bound completes the proof. □

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 $\pi_t = \mathcal{N}\left(\mathbf{w}_t, \sigma^2 \mathbf{I}_d\right)$ where $\mathbf{w}_t$ is the weight vector obtained after the t^{th} epoch. We assume that the priors are obtained from the learning algorithm being applied to the sample S_{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 $A : \mathcal{Z}^m \to \mathcal{M}(\mathcal{W})$, for any loss $l : \mathcal{W} \times \mathcal{Z} \to [0,1]$ and for any $\delta > 0$ then for any $\pi_t \in \mathbf{P}$ we have that*

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

Proof. We can apply Theorem 5.2 with $\phi(\mathbf{w}, S) = \exp\left(\frac{\alpha-1}{\alpha} m\text{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right)\right)$ and $\alpha = 2$. To deduce that for all $\pi_t \in \mathbf{P}$ we have

$$\mathbb{P}_{S\sim\mathcal{D}^m, \mathbf{w}\sim\rho_S}\Bigg(\text{kl}\left(\hat{R}(\mathbf{w}), R(\mathbf{w})\right) \\ \leq \frac{1}{m}\left(D_2(\rho_S, \pi_t) + \log\left(\frac{8T}{\delta^3}\mathbb{E}_{S'\sim\mathcal{D}^m}\mathbb{E}_{\mathbf{w}'\sim\pi_t}\left(\exp\left(m\text{kl}\left(\hat{R}(\mathbf{w}'), R(\mathbf{w})\right)\right)\right)\right)\right) \geq 1-\delta.$$

Note that the empirical risk in the exponential is with respect to the distribution S' where as the empirical risk on the left hand side of the inequality is with respect to S. Recall, the upper bound we determined in the proof of Theorem 3.12,

$$\mathbb{E}_{S'\sim\mathcal{D}^m}\mathbb{E}_{\mathbf{w}'\sim\pi_t}\left(\exp\left(m\text{kl}\left(\hat{R}(\mathbf{w}'), R(\mathbf{w})\right)\right)\right) \leq 2\sqrt{m}.$$

Furthermore, it is known that for $\rho_S = \mathcal{N}(\mathbf{w}, \sigma^2 I_d)$ and $\pi_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 I_d)$ that

$$D_2(\rho_S, \pi_t) = \frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2}.$$

Putting this and our bound into our deductions from Theorem 5.2 completes the proof of the corollary. □

Corollary 5.4 ([24])**.** *Under the assumptions of Corollary 5.3 with $\delta \in (0,1)$ and for all $\pi_t \in \mathbf{P}$ we have that*

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

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 $\epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_d\right)$ 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.

The proof of these individual statements follow the same structure. We will only prove the last of these with the aid of Theorem 5.4.3 proven below. The first two can be proven in a similar way using Theorem 1 (i) from [22] and Proposition 3.1 from [6] respectively.

Lemma 5.4.1. *For $\rho_S = \mathcal{N}\left(\mathbf{w}, \sigma^2 I_d\right)$ and $\pi = \mathcal{N}\left(\mathbf{v}, \sigma^2 I_d\right)$, we have that*

$$\log\left(\frac{\rho_S(\mathbf{w}+\epsilon)}{\pi(\mathbf{w}+\epsilon)}\right) = \frac{1}{2\sigma^2}\left(\|\mathbf{w}+\epsilon-\mathbf{v}\|_2^2 - \|\epsilon\|_2^2\right),$$

where $\epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I_d\right)$ such that $\mathbf{w}+\epsilon$ acts as the weights sampled from $\mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$.

Proof. This follows from simple computations after recalling that

$$\rho_S(\mathbf{w}+\epsilon) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^d \exp\left(-\frac{1}{2\sigma^2}\|\epsilon\|_2^2\right), \text{ and } \pi(\mathbf{w}+\epsilon) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^d \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{w}+\epsilon-\mathbf{v}\|_2^2\right).$$

So this completes the proof of the lemma. □ ■

Lemma 5.4.2 ([7])**.** *For any positive λ and $\mathbf{w} \in \mathcal{W}$, we have that*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(\Phi_{\frac{\lambda}{m}}(R(\mathbf{w})) - \hat{R}(\mathbf{w})\right)\right)\right) \leq 1.$$

Proof. Define the Bernoulli random variables $\sigma_i(\mathbf{w}) = \mathbb{I}\left(h_{\mathbf{w}}(x_i) \neq y_i\right)$. Using independence, the concavity of $\log$ and $\lambda > 0$ we deduce that

$$\begin{aligned}\log\left(\mathbb{E}\left(\exp\left(-\lambda\hat{R}(\mathbf{w})\right)\right)\right) &= \sum_{i=1}^m \log\left(\mathbb{E}\left(\exp\left(-\frac{\lambda}{m}\sigma_i\right)\right)\right) \\ &\leq m\log\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left(\exp\left(-\frac{\lambda}{m}\sigma_i\right)\right)\right).\end{aligned}$$

Now note that

$$R(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}(\sigma_i)$$

and becaus the σ_i are Bernoulli random variables we have that

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left(\exp\left(-\frac{\lambda}{m}\sigma_i\right)\right) &= \frac{1}{m}\sum_{i=1}^m \left((1-\mathbb{E}(\sigma_i)) + \exp\left(-\frac{\lambda}{m}\right)\mathbb{E}(\sigma_i)\right) \\ &= \frac{1}{m}\sum_{i=1}^m \left(\mathbb{E}(\sigma_i)\left(\exp\left(-\frac{\lambda}{m}\right)-1\right)+1\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\Phi_{\frac{\lambda}{m}}(R(\mathbf{w})) &= \frac{1}{-\lambda} m\log\left(1-\left(1-\exp\left(-\frac{\lambda}{m}\right)\right)\frac{1}{m}\sum_{i=1}^m \mathbb{E}(\sigma_i)\right) \\ &= \frac{1}{-\lambda} m\log\left(\frac{1}{m}\sum_{i=1}^m \left(\mathbb{E}(\sigma_i)\left(\exp\left(-\frac{\lambda}{m}\right)-1\right)+1\right)\right) \\ &= \frac{1}{-\lambda} m\log\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left(\exp\left(-\frac{\lambda}{m}\sigma_i\right)\right)\right).\end{aligned}$$

From which we conclude that

$$\log\left(\mathbb{E}\left(\exp\left(-\lambda\hat{R}(\mathbf{w})\right)\right)\right) \leq -\lambda\Phi_{\frac{\lambda}{m}}(R(\mathbf{w})).$$

Re-arranging the terms completes the proof of the lemma. ■

Theorem 5.4.3 ([7]). *For any positive λ, any posterior distribution $\rho \in \mathcal{M}(\mathcal{W})$, then*

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

Proof. To prove this we start from Lemma 5.4.2 and integrate with respect to π to get that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(\Phi_\lambda m\left(R(\pi)\right) - \hat{R}(\pi)\right)\right)\right) \leq 1.$$

Which for any posterior ρ can be written as

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(\Phi_\lambda m\left(R(\rho)\right) - \hat{R}(\rho)\right) - \log\left(\frac{d\rho}{d\pi}\right) + \log(\delta)\right)\right) \leq \delta.$$

From this we can deduce using by Markov's inequality that

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\rho) \geq \Phi^{-1}_{\frac{\lambda}{m}}\left(\hat{R}(\rho) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\frac{d\rho}{d\pi}\right)\right)\right) \\
&\qquad = \mathbb{P}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(\Phi_\lambda m\left(R(\rho)\right) - \hat{R}(\rho)\right) - \log\left(\frac{d\rho}{d\pi}\right) + \log(\delta)\right) \geq e^0\right) \\
&\qquad \leq \mathbb{E}_{S\sim\mathcal{D}^m}\left(\exp\left(\lambda\left(\Phi_\lambda m\left(R(\rho)\right) - \hat{R}(\rho)\right) - \log\left(\frac{d\rho}{d\pi}\right) + \log(\delta)\right)\right) \\
&\qquad \leq \delta
\end{aligned}$$

from which when we take complements we complete the proof of the theorem. ■

Proof. Apply Theorem 5.4.3 $T|\mathbf{C}|$ times with confidence $\frac{\delta}{T|\mathbf{C}|}$. For each prior $\pi_t \in \mathbf{P}$ and hyperparameter $c \in \mathbf{C}$, we have that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\rho_S) \leq \frac{1}{1-e^{-c}}\left(1 - \exp\left(-c\hat{R}(\rho_S) - \frac{1}{m}\left(\log\left(\frac{\rho_S(\mathbf{w})}{\pi_t(\mathbf{w})}\right) + \log\left(\frac{T|\mathbf{C}|}{\delta}\right)\right)\right)\right)\right) \geq 1 - \frac{\delta}{T|\mathbf{C}|}.$$

Applying a union bound argument and Lemma 5.4.1 the conclusions of the theorem follows which completes the proof. □

5.2 PAC-Bayes Compression Bounds

We will now see how compression ideas can be capitalized to tighten PAC-Bayes bounds. The work of [19] 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 [19] utilizes a refined version of Theorem 3.11.

Theorem 5.5 ([7]). *Let L be a 0-1 valued loss function. Let π be a probability measure on the parameter space, and let $\alpha > 1, \delta > 0$. Then,*

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

Proof. We start from Theorem 3.11 and try to optimize the bound with respect to λ. Let us introduce the parameter $\alpha > 1$ and let $\Lambda = \left\{\alpha^k : k \in \mathbb{N}\right\}$ on which we define the probability measure $\nu\left(\alpha^k\right) = \frac{1}{(k+1)(k+2)}$. Now for each $k \in \mathbb{N}$ apply Theorem 3.11 with $\lambda = \alpha^k$ and confidence $1 - \frac{\delta}{(k+1)(k+2)}$. Now apply a union bound argument to conclude that

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

We note that $\lambda \in (1,\infty)$ (as for $\lambda < 1$ we get a bound larger than 1) and so there is a $\lambda' \in \Lambda$ such that

$$\frac{\lambda}{\alpha} \leq \lambda' \leq \lambda.$$

Moreover, for any $q \in (0,1)$ we have that $\beta \mapsto \Phi_\beta(q)$ is increasing on $\mathbb{R}_+$. Therefore,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(R(\rho) \leq \inf_{\lambda\in(1,\infty)} \Phi^{-1}_{\frac{\lambda}{m}}\left(\hat{R}(\rho) + \frac{\alpha}{\lambda}\left(\mathrm{KL}(\rho,\pi) - \log(\delta) + 2\log\left(\frac{\log\left(\alpha^2\lambda'\right)}{\log(\alpha)}\right)\right)\right)\right) \geq 1-\delta,$$

which completes the proof. □

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

Theorem 5.6 ([19])**.** *Let $|\mathbf{w}|_c$ denote the number of bits required to represent hypothesis $h_{\mathbf{w}}$ using some pre-specified coding c. Let ρ denote the point mass distribution at $\hat{\mathbf{w}}$ which is the compression of $\mathbf{w}$ and corresponds to the compressed model $h_{\hat{\mathbf{w}}}$. Let M denote any probability measure on the positive integers. Then there exists a prior π_c such that*

$$\mathrm{KL}(\rho,\pi_c) \leq |\hat{\mathbf{w}}|_c \log(2) - \log\left(M\left(|\hat{\mathbf{w}}|_c\right)\right).$$

Proof. Let $\mathcal{W}_c \subseteq \mathcal{W}$ be the set of compressed weights. Then let π_c be a distribution on $\mathcal{W}_c$ defined by

$$\pi_c(\mathbf{w}) = \frac{1}{Z} M(|\mathbf{w}|_c)\cdot 2^{-|\mathbf{w}|_c}, \text{ where } Z = \sum_{\mathbf{w}\in\mathcal{W}_c} M(|\mathbf{w}|_c)\cdot 2^{-|\mathbf{w}|_c}.$$

As c is injective on $\mathcal{W}_c$ we have that $Z \leq 1$. Therefore,

$$\begin{aligned}\mathrm{KL}(\rho,\pi_c) = \log\left(\frac{\rho\left(\hat{\mathbf{w}}\right)}{\pi_c\left(\hat{\mathbf{w}}\right)}\right)\rho\left(\hat{\mathbf{w}}\right) &= -\log\left(\pi_c\left(\hat{\mathbf{w}}\right)\right) \\ &= \log(Z) + |\hat{\mathbf{w}}|_c\log(2) - \log\left(M\left(|\hat{\mathbf{w}}|_c\right)\right) \\ &\leq |\hat{\mathbf{w}}|_c\log(2) - \log\left(M\left(|\hat{\mathbf{w}}|_c\right)\right).\end{aligned}$$

Which completes the proof of the theorem. □

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 hypotheses which is translated to the space $\mathcal{W}$. By exploiting structure in the hypothesis class the bound can be improved substantially.*

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

- $S = \{s_1, \ldots, s_k\} \subseteq \{1, \ldots, d\}$ is the location of the non-zero weights,

- $C = \{c_1, \ldots, c_r\} \subseteq \mathbb{R}$, is a codebook, and
- $Q = (q_1, \ldots, q_k)$ for $q_i \in \{1, \ldots, r\}$ 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 $\rho \sim \mathcal{N}\left(\mathbf{w}, \sigma^2 J\right)$, with J being a diagonal matrix.

Theorem 5.8 ([19]). *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 σ^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 π such that*

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

Proof. The following is a proof by construction, that is we construct prior π with the desired property. To do this we want to express the prior as a mixture over all possible compressions provided by the algorithm. We first define the mixture component

$$\pi_{S,Q,C} = \mathcal{N}\left(\mathbf{w}(S, Q, C), \tau^2\right).$$

We then define our prior to be a weighted mixture over all possible compressions, that is

$$\pi = \frac{1}{Z} \sum_{S,Q,C} M\left(|S|_c + |C|_c + k\lceil \log(r) \rceil\right) \cdot 2^{-|S|_c - |C|_c - k\lceil \log(r) \rceil} \pi_{S,Q,C}.$$

Where $Z \leq 1$ as the compression scheme is injective. Let $\left(\hat{S}, \hat{Q}, \hat{C}\right)$ be the output of our compression algorithm, so that out posterior ρ is $\mathcal{N}\left(\mathbf{w}\left(\hat{S}, \hat{Q}, \hat{C}\right), \sigma^2\right)$. Therefore,

$$\begin{aligned} \mathrm{KL}(\rho, \pi) &\leq \mathrm{KL}\left(\rho, \sum_{S,Q,C} M\left(|S|_c + |C|_c + k\lceil \log(r) \rceil\right) \cdot 2^{-|S|_c - |C|_c - k\lceil \log(r) \rceil} \pi_{S,Q,C} \right) \\ &\leq \mathrm{KL}\left(\rho, \sum_{Q} M\left(\left|\hat{S}\right|_c + \left|\hat{C}\right|_c + k\lceil \log\left(\hat{r}\right) \rceil \right) \cdot 2^{-\left|\hat{S}\right|_c - \left|\hat{C}\right|_c - k\lceil \log(\hat{r}) \rceil} \pi_{\hat{S},Q,\hat{C}} \right) \\ &\leq \left(\left|\hat{S}\right|_c + \left|\hat{C}\right|_c + k\lceil \log\left(\hat{r}\right) \rceil \right) \log(2) + \log\left(M\left(\left|\hat{S}\right|_c + \left|\hat{C}\right|_c + k\lceil \log\left(\hat{r}\right) \rceil \right) \right) + \mathrm{KL}\left(\rho, \sum_{Q} \pi_{\hat{S},Q,\hat{C}} \right) \end{aligned}$$

Let $\phi_\tau = \mathcal{N}\left(0, \tau^2\right)$. Then as the mixture term is independent across coordinates we have that

$$\left(\sum_{Q} \pi_{\hat{S},Q,\hat{C}} \right)(x) = \sum_{q^1, \ldots q^k = 1}^{r} \prod_{i=1}^{k} \phi_\tau\left(x_i - \hat{c}_{q^i}\right) = \prod_{i=1}^{k} \sum_{q^i=1}^{r} \phi_\tau\left(x_i - \hat{c}_{q^i}\right).$$

Furthermore, as ρ is independent over the coordinates, we get that

$$\mathrm{KL}\left(\rho, \sum_{Q} \pi_{\hat{S},Q,\hat{C}} k\right) = \sum_{i=1}^{k} \mathrm{KL}\left(\rho_i, \sum_{q^i=1}^{r} \mathcal{N}\left(\hat{c}_{q^i}, \tau^2\right)\right),$$

from which the result follows and completes the proof of the theorem. □

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.

6 Appendix

6.1 Extensions to Convolutional Neural Networks

In this section, we extend the ideas of Section 2.3 to convolutional neural networks (CNN) [17]. This extension is not trivial due to the parameter sharing that occurs in the CNN architecture. To investigate these ideas we update our notation from that of Section 2.3. In particular, we suppose that the i^{th} layer has an image dimension of $n_1^i \times n_2^i$, where each pixel has l^i channels, and the filter at layer i has size $\kappa_i \times \kappa_i$ with stride s_i. The convolutional filter has dimension $l^{i-1} \times l^i \times \kappa_i \times \kappa_i$. If we apply Algorithm 3 to each copy of the filter then the number of new parameters grows proportionally to $n_1^i n_2^i$, which is undesirable. On the other hand, compressing the filter once and re-using it for all patches removes the implicit assumption that the noise generated by the compression behaves similar to a Gaussian as the shared filters introduces correlations. To solve these issues Algorithm 8 generates p-wise independent compressed filters for each convolution location. This results in p more parameters than a single compression, but if p grows logarithmically with respect to the relevant parameters then the filters behave like fully independent filters. To proceed with this idea we need to introduce some operations. For $k' \leq k$ let Y be a k^{th} order tensor and Z a $(k')^{\text{th}}$ order tensor with a matching dimensionality to the last k'-dimensions of Y. The product operator $\times_{k'}$ when given tensors Y and Z returns a $(k-k')^{\text{th}}$ order tensor as follows

$$(Y \times_{k'} Z)_{i_1,\ldots,i_{k-k'}} = \left\langle Y_{i_1,\ldots,i_{k-k'}}, Z \right\rangle = \left\langle \operatorname{vec}\left(Y_{i_1,\ldots,i_{k-k'}}\right), \operatorname{vec}(Z) \right\rangle.$$

Let $X \in \mathbb{R}^{l \times n_1 \times n_2}$ be an $n \times n$ image where the pixels have l features. Denote the $\kappa \times \kappa$ sub-image starting from pixel (i,j) by $X_{(i,j),\kappa} \in \mathbb{R}^{l \times \kappa \times \kappa}$. Let $A \in \mathbb{R}^{l' \times l \times \kappa \times \kappa}$ be a convolutional weight tensor. The convolutional operator with stride s can then be defined as

$$(A *_s X)_{i,j} = A \times_3 X_{(s(i-1)+1,s(j+1)+1),\kappa}$$

for $1 \leq i \leq \left\lfloor \frac{n_1-\kappa}{s} \right\rfloor =: n_1'$ and $1 \leq j \leq \left\lfloor \frac{n_2-\kappa}{s} \right\rfloor =: n_2'$ so that $A *_s X \in \mathbb{R}^{l' \times n_1' \times n_2'}$. Algorithm 8 generates p-wise independent filters $\hat{A}_{(a,b)}$ for each convolution location $(a,b) \in [n_1'] \times [n_2']$ and so $\hat{A} *_s X$ will be used to denote the convolution operator

$$\left(\left(\hat{A} *_s X\right)_{i,j}\right) = \hat{A}_{(i,j)} \times_3 X_{(s(i-1)+1,s(j+1)+1),\kappa}$$

for $1 \leq i \leq n_1'$ and $1 \leq j \leq n_2'$. With this we see that for any $i > 1$ we have

$$x^{i+1} = \phi\left(A^i *_{s_i} x^i\right), \text{ and } x^j = M^{ij}\left(x^i\right) = J^{ij}_{x^i} \times_3 x^i.$$

Definition 6.1. *For any two layer $i \leq j$, we define the inter-layer cushion $\mu_{i,j}$ as the largest number such that for any $(x,y) \in S$ we have that*

$$\mu_{i,j} \frac{1}{\sqrt{n_1^i n_2^i}} \left\| J^{i,j}_{x^i} \right\|_F \left\| x^i \right\| \leq \left\| J^{i,j}_{x^i} x^i \right\|.$$

For any layer i let the minimal inter-layer cushion be $\mu_{i\to} = \min_{i \leq j \leq d} \mu_{i,j} = \min\left(\frac{1}{\sqrt{l^i}}, \min_{i<j\leq d} \mu_{i,j}\right)$.

Definition 6.2. *Let $J^{i,j}_x \in \mathbb{R}^{l^i \times n_1^i \times n_2^i \times l^j \times n_1^j \times n_2^j}$ be the Jacobian of $M^{i,j}$ at x. We say that the Jacobian is β well-distributed if for any $(x,y) \in S$, any i,j and any $(a,b) \in \left[n_1^i \times n_2^i\right]$ we have that*

$$\left\| \left[J^{i,j}_x\right]_{:,a,b,:,:,:} \right\|_F \leq \frac{\beta}{\sqrt{n_1^i n_2^i}} \left\| J^{i,j}_x \right\|_F.$$

For any $\delta > 0$, $\epsilon \leq 1$, let $G = \left\{\left(U^i, V^i\right)\right\}_{i=1}^m$ be a set of matrix/vector pairs where $U \in \mathbb{R}^{l' \times n_1' \times n_2' \times n_u}$ and $V \in \mathbb{R}^{l \times n_1 \times n_2}$, let $\hat{A}_{(i,j)} \in \mathbb{R}^{l \times l'}$ be the output of Algorithm 8 with $\eta = \frac{\delta}{\eta}$ and $\Delta_{(i,j)} = \hat{A}_{(i,j)} - A$. Suppose the U's are β-well-distributed. Then for any $(U,V) \in G$ we have that

$$\mathbb{P}\left(\|U \times_3 (\Delta *_s V)\| \leq \frac{\eta\beta}{\sqrt{l_1' l_2'}} \|A\|_F \|U\|_F \|V\|_F \right) \geq 1 - \delta \quad (\star). \tag{5}$$

Algorithm 8 $(A, \epsilon, \eta, n_1' \times n_2'))$

Require: Convolution Tensor $A \in \mathbb{R}^{l' \times l \times \kappa \times \kappa}$, error parameters ϵ, η.
Ensure: Generate $n_1' \times n_2'$ different tensors $\hat{A}_{(i,j)}$ $((i,j) \in [n_1'] \times [n_2'])$ that satisfy (5).

Let $k = \frac{Q\lceil \frac{\kappa}{s} \rceil^2 \left(\log\left(\frac{1}{\eta}\right)\right)^2}{\epsilon^2}$ for a large enough universal constant Q.
Let $p = \log\left(\frac{1}{\eta}\right)$.
Sample a uniformly random subspace $\mathcal{S}$ of $l' \times l \times \kappa \times \kappa$ of dimension $k \times p$.
for $(i,j) \in [n_1'] \times [n_2']$ **do**
  Sample k matrices $M_1, \ldots, M_k \in \mathcal{N}(0,1)^{l' \times l \times \kappa \times \kappa}$ with random i.i.d entries.
  for $k' = 1 \to k$ **do**
    Let $M_{k'}' = \sqrt{\frac{ll'\kappa^2}{kp}} \mathrm{Proj}_{\mathcal{S}}(M_{k'})$.
    Let $Z_{k'} = \langle A, M_{k'}' \rangle M_{k'}'$.
  end for
  Let $\hat{A}_{(i,j)} = \frac{1}{k} \sum_{k'=1}^{k} Z_{k'}$.
end for

Algorithm 8 is designed to generate different compressed filters $\hat{A}_{i,j}$ in a way that keeps the total number of parameters small, but also ensures that the $\hat{A}_{i,j}$'s behave similarly to the compressed filters that would generated if Algorithm 3 were applied to each location independently.

Theorem 6.3. *For any convolutional neural network $h_{\mathbf{w}}$ with $\rho_\delta \geq 3d$, any probability $0 < \delta \leq 1$ and any margin γ, then Algorithm 8 generates weights $\tilde{\mathbf{w}}$ for the network $h_{\tilde{\mathbf{w}}}$ such that*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_0(h_{\tilde{\mathbf{w}}}) \leq \hat{L}_\gamma(h_{\mathbf{w}}) + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{(x,y)\in S} \|h_{\mathbf{w}}(x)\|_2^2 \sum_{i=1}^{d} \frac{\beta^2 \left(\left\lceil \frac{\kappa_i}{s_i} \right\rceil\right)^2}{\mu_i^2 \mu_{i\to}^2}}{\gamma^2 m}} \right) \right) \geq 1 - \delta,$$

where $\mu_i, \mu_{i\to}, c, \rho_\delta$ and β are layer cushion, inter-layer cushion, activation contraction, inter-layer smoothness and well-distributed Jacobian respectively. Furthermore, κ_i and s_i are the filter and stride in layer i.

Corollary 6.4. *For any convolutional neural network $h_{\mathbf{w}}$ with $\rho_\delta \geq 3d$, any probability $0 < \delta \leq 1$ and any margin γ, then Algorithm 8 generates weights $\tilde{\mathbf{w}}$ for the network $h_{\tilde{\mathbf{w}}}$ such that*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_0(h_{\tilde{\mathbf{w}}}) \leq \hat{L}_\gamma(h_{\mathbf{w}}) + \zeta + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{(x,y)\in S} \|h_{\mathbf{w}}(x)\|_2^2 \sum_{i=1}^{d} \frac{\beta^2 \left(\left\lceil \frac{\kappa_i}{s_i} \right\rceil\right)^2}{\mu_i^2 \mu_{i\to}^2}}{\gamma^2 m}} \right) \right) \geq 1 - \delta,$$

where $\mu_i, \mu_{i\to}, c, \rho_\delta$ and β are layer cushion, inter-layer cushion, activation contraction, inter-layer smoothness and well-distributed Jacobian respectively measured on a $1 - \zeta$ fraction of the training set S. Furthermore, κ_i and s_i are the filter and stride in layer i.

6.2 Current State of the Art PAC-Bayes Bounds

We have seen that PAC-Bayes bounds provide a theoretical perspective on the learning process and the consequences it has on the performance of the learned classifier. In practice, we would ideally want these bounds to be meaningful. When implemented naively they produce vacuous bounds that provide no information. The first implementation of non-vacuous PAC-Bayes was discussed in Section 3.2 with the work [15] that focused on a particular setting to get the non-vacuous bounds. Since then there have been directed efforts to improve the tightness of these bounds and extend the success to different contexts. Currently, the tightest bounds seen in practice come from the work of [25]. In this section, we will discuss the work and see how it is a development of some previous work we have discussed. The work of [25] is an extension of the work

of [19] and follows the same compression paradigm that was first considered by [17]. In [25] the tighter generalization bounds are achieved by first restricting to lower-dimensional settings using a notion called intrinsic dimensionality. Then they develop more aggressive quantization schemes that are adapted to the problem at hand.

6.2.1 The PAC-Bayes Foundations

Throughout this section, we will adopt the same notation as the rest of this report. Consider Theorem 2.1, the $\log(M)$ term counts the number of bits needed to specify any hypothesis $h_{\mathbf{w}}$ with $\mathbf{w} \in \mathcal{W}$, supposing that we assume each hypothesis is equally likely. If instead we have some prior belief on the likely hypotheses we can construct a variable length code that uses fewer bits to specify those hypotheses. For a prior distribution π, then for any $\mathbf{w} \in \mathcal{W}$ the number of bits required to represent hypothesis $h_{\mathbf{w}}$ is $\log_2\left(\frac{1}{\pi(\mathbf{w})}\right)$ when using an optimal compression code for π. Furthermore, if we consider a set of good distributions ρ and we do not care which element of Q we arrive at, we can gain some bits back. In particular, the average number of bits to code a sample from ρ using π is the cross entropy $H(\rho, \pi)$ and since we are agnostic to the sample from ρ we get back $H(\rho)$ bits. Therefore, the average number of bits is

$$H(\rho, \pi) - H(\rho) = \mathrm{KL}(\rho, \pi).$$

Definition 6.5. *For probability measures ρ and π on a state space $\mathcal{X}$ that are absolutely continuous with respect to some measure λ, then*

$$H(\rho, \pi) = \int_{\mathcal{X}} \rho(x) \log\left(\pi(x)\right) d\lambda(x),$$

where $H(\rho) := H(\rho, \rho)$.

With these improvements, we can get bounds such as Theorem 3.10. For this work, we will work with Theorem 5.5 to get the generalization bounds. The prior that we will use here is known as the universal prior and explicitly penalizes the minimum compressed length of the hypothesis,

$$\pi(\mathbf{w}) = \frac{2^{-K(\mathbf{w})}}{Z}.$$

Then using a point mass posterior on the single parameter $\mathbf{w}^*$ we get that

$$\mathrm{KL}\left(\mathbf{I}_{\{\mathbf{w}=\mathbf{w}^*\}}, \pi\right) = \log\left(\frac{1}{\pi(\mathbf{w}^*)}\right) \leq K\left(\mathbf{w}^*\right)\log(2) \leq l(\mathbf{w}^*)\log(2) + 2\log\left(l(\mathbf{w}^*)\right),$$

where $l(\mathbf{w})$ is the length of the program that reproduces $\mathbf{w}$. Improving the tightness of our bounds comes about by reducing the compressed length $l(\mathbf{w}^*)$ for the $\mathbf{w}^*$ achieved through training. For this work, the method for model compression consists of two components. One component is reducing the dimensionality of the problem, and the second is developing an aggressive quantization scheme.

6.2.2 Finding Random Subspaces

A neural network parameterized by the weight vector $\mathbf{w} \in \mathbb{R}^D$ is often optimized through gradient descent so that the updates occur in the D-dimensional loss landscape. Despite D being very large the optimization process is relatively stable and converges to simple solutions. However, we can work in a reduced dimension $d < D$ (referred to as the intrinsic dimension) by considering

$$\mathbf{w} = \mathbf{w}_0 + P\hat{\mathbf{w}},$$

where $\mathbf{w}_0 \in \mathbb{R}^D$ is the initialized weight, $P \in \mathbb{R}^{D\times d}$ is such that $P^\top P \approx I_{d\times d}$ and $\hat{\mathbf{w}} \in \mathbb{R}^d$. Now the vector $\hat{\mathbf{w}}$ is optimized so that the updates take place on a d-dimensional landscape. Finding the smallest value of d for which we still attain good performance on the problem at hand is the bottleneck to this approach. The work lies in finding projection P that is stable under training and finding optimal subspaces in which to optimize in. Imposing the condition $P^\top P \approx I_{d\times d}$ solves the first concern, for the next, we consider three possible methods for constructing P.

1. Kronecker Sum Projector: Construct the matrix

$$P_{\oplus} = \frac{\mathbf{1} \otimes R_1 + R_2 \otimes \mathbf{1}}{\sqrt{2D}}$$

where $\otimes$ is the Kronecker product, $R_1, R_2 \sim \mathcal{N}(0,1)^{\sqrt{D} \times d}$. Note that $P_{\oplus}^{\top} P_{\oplus} = I_{d \times d} + O\left(\frac{1}{\sqrt{D}}\right)$. Applying this to a vector $\hat{\mathbf{w}} \in \mathbb{R}^d$ takes $O\left(d\sqrt{D}\right)$ time.

2. Kronecker Product Projector: Construct the matrix

$$P_{\otimes} = \frac{Q_1 \otimes Q_2}{\sqrt{D}}$$

where $Q_1, Q_2 \sim \mathcal{N}(0,1)^{\sqrt{D} \times \sqrt{d}}$. Note that $P_{\otimes}^{\top} P_{\otimes} = I_{d \times d} + O\left(\frac{1}{D^{\frac{1}{4}}}\right)$. Applying this to a vector $\hat{\mathbf{w}} \in \mathbb{R}^d$ takes $O\left(\sqrt{dD}\right)$ time.

6.2.3 Quantization and Training

For the full precision weight vector $\mathbf{w} = (w_1, \ldots, w_d) \in \mathbb{R}^d$ and vector $c = (c_1, \ldots, c_L) \in \mathbb{R}^L$ of L quantization levels, construct the quantized vector $\tilde{\mathbf{w}} \in \mathbb{R}^d$ where $\tilde{w}_i = c_{q(i)}$ where $q(i) = \text{argmin}_k |w_i - c_k|$. The vector c is learned alongside $\mathbf{w}$ where the gradients are defined as

$$\frac{\partial \tilde{w}_i}{\partial w_j} = \delta_{ij}, \text{ and } \frac{\partial \tilde{w}_i}{\partial c_k} = \mathbf{I}_{q(i)=k}.$$

c is initialized to have uniform spacing between the minimum and maximum values of $\mathbf{w}$, or using k-means. The latter approach refers to a quantization scheme proposed in [14] where for $k = L$ we partition the weights into clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ with $c_1, \ldots, c_k$ such that

$$\text{argmin}_{\mathcal{C}_1, \ldots, \mathcal{C}_k} \left(\sum_{i=1}^{k} \sum_{w \in \mathcal{C}_i} |w_i - c_i|^2 \right), \quad \text{for } c_i = \frac{1}{|\mathcal{C}_i|} \sum_{w \in \mathcal{C}_i} w.$$

Next, we capitalize on the fact that certain quantization levels will be more likely than others to introduce a variable length coding scheme. For each level c_k associate the probability p_k and apply arithmetic coding. Each arithmetic coding of $\mathbf{w}$ takes at most $\lceil d \times H(p) \rceil$ bits, where p is the discrete distribution of the p_k's. Considering the total number of bits we see that

$$l(\mathbf{w}) \leq \lceil d \times H(p) \rceil + L \times (16 + \lceil \log_2(d) \rceil) + 2$$

as $L \times \lceil \log_2(d) \rceil$ bits are required for the probabilities p_k and $16L$ bites for the codebook c.

6.2.4 Optimization

Note that the smaller the intrinsic dimension d is the closer that our trained weight will be to the initialized weight $\mathbf{w}_0$. Therefore, $\mathbf{w}_0$ is more likely under our universal prior. Recall, that we must therefore condition on $\mathbf{w}_0$ to generate our prior. Similarly, if we optimize over different hyper-parameters such as d, L or the learning rate (η), we must encode these into the prior and pay a penalty for optimizing over them. To do this we simply redefine our weight vector to be $\mathbf{w}' = (\mathbf{w}, d, L, \eta)$ and so our prior becomes

$$\pi(\mathbf{w}') = \frac{2^{-K(\mathbf{w}')}}{Z},$$

where now we have that

$$K(\mathbf{w}') \leq K(\mathbf{w}|d, L) + K(d) + K(L) + K(\eta).$$

Typically, we optimize these hyper-parameters over finite sets and so we can bound these terms by the ceiling of $\log_2$ of the size of these sets. The process we have discussed can be summarized in Algorithm 9.

Algorithm 9 Compute PAC-Bayes Compression Bound

Require: Neural network $h_{\mathbf{w}}$, training data set $S = \{(x_i, y_i)\}_{i=1}^m$, Clusters L, Intrinsic Dimension d, Confidence $1-\delta$, Prior distribution π

function COMPUTEBOUND($h_{\mathbf{w}}, L, d, S, \delta, \pi$)
  $\mathbf{w} \leftarrow \text{TRAINID}(h_{\mathbf{w}}, d, S)$
  $\tilde{\mathbf{w}} \leftarrow \text{TRAINQUANTIZE}\left(h_{\mathbf{w}}, L, S\right).$
  Compute quantized empirical risk $\hat{R}(\tilde{\mathbf{w}})$.
  $\text{KL}(\rho, \pi) \leftarrow \text{GETKL}\left(\tilde{w}, \pi\right).$
  return $\text{GETCATONIBOUND}\left(\hat{R}(\tilde{w}), \text{KL}(\rho, \pi), \delta, m\right)$
end function
function TRAINQUANTIZE($\mathbf{w}, L, S$)
  Initialize $c \leftarrow \text{GETCLUSTERS}\left(\mathbf{w}, L\right).$
  for $i = 1 \rightarrow \text{quantepochs}$ **do**
    $\begin{pmatrix} c \\ \mathbf{w} \end{pmatrix} \leftarrow \begin{pmatrix} c - \eta \nabla_c \mathcal{L}\left(\mathbf{w}, c\right) \\ \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, c) \end{pmatrix}.$
  end for
  return $\tilde{w}$
end function
function GETKL($\tilde{\mathbf{w}}, \pi$)
  $c, \text{count} \leftarrow \text{GETUNIQUEVALSCOUNTS}\left(\tilde{\mathbf{w}}\right).$
  $\text{messagesize} \leftarrow \text{DOARITHMETICENCODING}\left(\tilde{w}, c, \text{count}\right).$
  $\text{messagesize} \leftarrow \text{messagesize} + \text{hyperparamsearch}$
  return $\text{messagesize} + 2 \times \log\left(\text{messagesize}\right).$
end function

6.3 Rademacher Complexity

Recall, that we have the space $\mathcal{Z}$ on which a distribution $\mathcal{D}$ is defined from which we draw an i.i.d sampled $S = \{(x_i, y_i)\}_{i=1}^m$. Suppose we have a class of functions $\mathcal{F} = \{f : \mathcal{Z} \rightarrow \mathbb{R}\}$.

Definition 6.6 ([9]). *The empirical Rademacher complexity of $\mathcal{F}$ is*

$$\hat{\mathfrak{R}}(\mathcal{F}) = \mathbb{E}_{\sigma \in \{\pm 1\}} \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f((x_i, y_i)) \right) \right),$$

where each σ_i is an independent random variable uniformly distribution on $\{\pm 1\}$.

Definition 6.7 ([9]). *The Rademacher complexity of $\mathcal{F}$ is*

$$\mathfrak{R}(\mathcal{F}) = \mathbb{E}_{S \sim \mathcal{D}^m} \left(\hat{\mathfrak{R}}(\mathcal{F}) \right).$$

Theorem 6.8 ([9]). *For a parameter $\delta \in (0,1)$ if $\mathcal{F} \subseteq \{f : \mathcal{Z} \rightarrow [0,1]\}$ then*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left(\mathbb{E}_{z \sim \mathcal{D}} \left(f(z) \right) \leq \frac{1}{m} \sum_{i=1}^m f(z_i) + 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{m}} \right) \geq 1 - \delta,$$

and

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left(\mathbb{E}_{z \sim \mathcal{D}} \left(f(z) \right) \leq \frac{1}{m} \sum_{i=1}^m f(z_i) + 2\hat{\mathfrak{R}}(\mathcal{F}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{m}} \right) \geq 1 - \delta.$$

Theorem 6.8.1 (McDiamrid Inequality). *Let $x_1, \ldots, x_n$ be independent random variables taking values in a set A and let $c_1, \ldots, c_n$ be positive real constants. If $\phi : A^n \rightarrow \mathbb{R}$ satisfies*

$$\sup_{x_1, \ldots, x_n, x_i'} \left| \phi(x_1, \ldots, x_i, \ldots, x_n) - \phi\left(x_1, \ldots, x_i', \ldots, x_n\right) \right| \leq c_i,$$

for $1 \leq i \leq n$*, then*

$$\mathbb{P}\left(\phi(x_1,\ldots,x_n) - \mathbb{E}\left(\phi(x_1,\ldots,x_n)\right) \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon}{\sum_{i=1}^{n} c_i^2}\right).$$

Proof. For a proof of this theorem refer to [13]. ■

Lemma 6.8.2. *The function*

$$\phi(S) = \sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h(x,y)\right) - \frac{1}{m}\sum_{i=1}^{m} h(x_i,y_i)\right)$$

satisfies

$$\sup_{z_1,\ldots,z_n,z_i'\in\mathcal{Z}} |\phi(z_1,\ldots,z_i,\ldots,z_m) - \phi(z_1,\ldots,z_i',\ldots,z_m)| \leq \frac{1}{m}.$$

Proof. Let $S = \{z_1,\ldots,z_m\}$ and $S' = \{z_1,\ldots,z_i',\ldots,z_m\}$ then

$$\begin{aligned}|\phi(S) - \phi\left(S'\right)| = \Bigg|&\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h(x,y)\right) - \frac{1}{m}\sum_{(x_j,y_j)\in S} h(x_j,y_j)\right)\\ &- \sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h(x,y)\right) - \frac{1}{m}\sum_{(x_j,y_j)\in S'} h(x_j,y_j)\right)\Bigg|.\end{aligned}$$

Let $h^* \in \mathcal{F}$ be the function the maximizes the supremum of $\phi(S)$, then

$$\begin{aligned}|\phi(S) - \phi\left(S'\right)| = \Bigg|&\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h^*(x,y)\right) - \frac{1}{m}\sum_{(x_j,y_j)\in S} h^*(x_j,y_j)\\ &- \sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h(x,y)\right) - \frac{1}{m}\sum_{(x_j,y_j)\in S'} h(x_j,y_j)\right)\Bigg|\end{aligned}$$

and because h^* can at best also maximize $\phi\left(S'\right)$ we also have that

$$\begin{aligned}|\phi(S) - \phi\left(S'\right)| &\leq \Bigg|\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h^*(x,y)\right) - \frac{1}{m}\sum_{(x_j,y_j)\in S} h^*(x_j,y_j)\\ &\qquad - \mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h^*(x,y)\right) - \frac{1}{m}\sum_{(x_j,y_j)\in S'} h^*(x_j,y_j)\Bigg|\\ &= \left|\frac{1}{m}\sum_{(x_j,y_j)\in S'} h^*(x_j,y_j) - \frac{1}{m}\sum_{(x_j,y_j)\in S} h^*(x_j,y_j)\right|.\end{aligned}$$

By using the definitions of S and S' this simplifies to

$$\begin{aligned}|\phi(S) - \phi\left(S'\right)| &\leq \frac{1}{m}\left|h^*(x_i,y_i) - h^*\left(x_i',y_i'\right)\right|\\ &\leq \frac{1}{m},\end{aligned}$$

which completes the proof of the lemma. ■

Proof. Lemma 6.8.2 shows that $\phi(S) = \sup_{h\in\mathcal{F}} \left(\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}\left(h(x,y)\right) - \frac{1}{m}\sum_{i=1}^m h(x_i,y_i)\right)$ satisfies the conditions of 6.8.1, therefore,

$$\mathbb{P}\left(\phi(S) - \mathbb{E}_{S'\sim\mathcal{D}^m}\left(\phi\left(S'\right)\right) \geq t\right) \leq \exp\left(-\frac{t^2}{m}\right).$$

With $t = \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{m}}$ we deduce that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\mathbb{E}_{\hat{S}\sim\mathcal{D}^m}(f(x,y)) \leq \frac{1}{m}\sum_{i=1}^m f(x_i,y_i) + \mathbb{E}_{\hat{S}'\sim\mathcal{D}^m}\left(\phi\left(\hat{S}'\right)\right)\right) \geq 1-\delta.$$

Now we need to bound the expectation of $\phi(S)$ using Rademacher complexity to complete the proof. Let $\tilde{S} = \{\tilde{z}_1,\ldots,\tilde{z}_m\}$ be a sample independent but identically distributed to S. As

$$\mathbb{E}_{\tilde{S}}\left(\frac{1}{m}\sum_{(x,y)\in\tilde{S}} h(x,y)\middle| S\right) = \mathbb{E}_{z\sim\mathcal{D}}\left(h(z)\right), \text{ and } \mathbb{E}_{\tilde{S}}\left(\frac{1}{m}\sum_{(x,y)\in S} h(x,y)\middle| S\right) = \frac{1}{m}\sum_{(x,y)\in S} h(x,y)$$

we deduce that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\phi(S)\right) = \mathbb{E}_{S\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\tilde{S}\sim\mathcal{D}^m}\left(\frac{1}{m}\sum_{(x,y)\in\tilde{S}}(h(x,y)) - \frac{1}{m}\sum_{(x,y)\in S} h(x,y)\middle| S\right)\right)\right).$$

We can apply Jensen's inequality as $\sup$ is convex to deduce that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\tilde{S}\sim\mathcal{D}^m}\left(\frac{1}{m}\sum_{(x,y)\in\tilde{S}} h(x,y) - \frac{1}{m}\sum_{(x,y)\in S} h(x,y)\middle| S\right)\right)\right)$$
$$\leq \mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{\tilde{S}\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{F}}\left(\frac{1}{m}\sum_{(x,y)\in\tilde{S}} h(x,y) - \frac{1}{m}\sum_{(x,y)\in S} h(x,y)\right)\right).$$

As $\mathbb{E}(\sigma_i) = 0$ we can multiply each term by σ_i, and in distribution we have $-\sigma_i = \sigma_i$ so that

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{\tilde{S}\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{F}}\left(\frac{1}{m}\sum_{(x,y)\in\tilde{S}}h(x,y)-\frac{1}{m}\sum_{(x,y)\in S}h(x,y)\right)\right)\\
&\qquad=\mathbb{E}_{\sigma\in\{\pm1\}^m}\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{\tilde{S}\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{F}}\left(\frac{1}{m}\sum_{(x,y)\in\tilde{S},\sigma_i\in\sigma}\sigma_i h(x,y)\right.\right.\\
&\qquad\quad\left.\left.-\frac{1}{m}\sum_{(x,y)\in S,\sigma_i\in\sigma}\sigma_i h(x,y)\right)\right)\\
&\qquad\le\mathbb{E}_{\sigma\in\{\pm1\}^m}\mathbb{E}_{S\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{F}}\left(\frac{1}{m}\sum_{(x,y)\in S,\sigma_i\in\sigma}\sigma_i h(x,y)\right)\right)\\
&\qquad\quad+\mathbb{E}_{\sigma\in\{\pm1\}^m}\mathbb{E}_{\tilde{S}\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{F}}\left(\frac{1}{m}\sum_{(x,y)\in\tilde{S},\sigma_i\in\sigma}\sigma_i h(x,y)\right)\right)\\
&\qquad=2\mathfrak{R}(\mathcal{F}),
\end{aligned}$$

which when substituted into our previous bounds completes the proof of the first statement. To obtain the second statement we note that $\hat{\mathfrak{R}}(\mathcal{F})$ satisfies Theorem 6.8.1 with constant $\frac{1}{m}$. Therefore, a second application of Theorem 6.8.1 with confidence level (where a confidence level of $\frac{\delta}{2}$ is used for each application) gives the desired result. □

If we let $\mathcal{F} = \{(x,y) \mapsto \mathbb{I}\left(h_{\mathbf{w}}(x)[y] \le \gamma + \max_{j\neq y} h_{\mathbf{w}}(x)[j]\right) : \mathbf{w} \in \mathcal{W}\}$ then for any $\delta \in (0,1)$ and $\mathbf{w} \in \mathcal{W}$ we have that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_\gamma\left(h_{\mathbf{w}}\right) \le \hat{L}_\gamma(h_{\mathbf{w}}) + 2\hat{\mathfrak{R}}(\mathcal{F}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{m}}\right) \ge 1-\delta.$$

Definition 6.9. *[26] Given a set $\mathcal{S}$ and a function $\rho : \mathcal{S}\times\mathcal{S} \to \mathbb{R}_+$, we call $(\mathcal{S},\rho)$ a pseudo-metric space if for all $x, y, z \in \mathcal{S}$ we have*

- $\rho(x,y) = \rho(y,x)$*,*
- $\rho(x,z) \le \rho(x,y) + \rho(y,z)$*, and*
- $\rho(x,x) = 0$*.*

Definition 6.10 ([26])**.** *Let $(\mathcal{S},\rho)$ be a pseudo-metric space and let $\epsilon > 0$. Then the set $\mathcal{C} \subseteq \int$ is an ϵ-cover of $(\mathcal{S},\rho)$ if for every $x \in \mathcal{S}$ there is a $y \in \mathcal{C}$ such that $\rho(x,y) \le \epsilon$. The set $\mathcal{C}$ is a minimal ϵ-cover if there is no other ϵ-cover with lower cardinality. The cardinality of any minimal ϵ-cover is the ϵ-covering number denoted $N(\mathcal{S},\rho,\epsilon)$.*

For a given training set $S = \{(x_i,y_i)\}_{i=1}^m$ we can consider the set

$$\mathcal{G} = \{(f(x_1,y_1),\ldots,f(x_m,y_m)) : f \in \mathcal{F}\}.$$

Theorem 6.11 ([21])**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{Z} \to [0,1]\}$ and $S \sim \mathcal{D}^m$ then*

$$\hat{\mathfrak{R}}(\mathcal{F}) \le \inf_{\epsilon>0}\left(\epsilon + \sqrt{\frac{2N(\mathcal{G},\rho,\epsilon)}{m}}\right).$$

Lemma 6.11.1 (Massart's Lemma[26])**.** *Let $\mathcal{T} \subseteq \mathbb{R}^n$ then we have that*

$$\mathfrak{R}(\mathcal{T}) \le \max_{t\in\mathcal{T}}\|t\|_2\frac{\sqrt{2\log|\mathcal{T}|}}{n}.$$

Proof ([12]). For all $a \geq 0$ we have that

$$\begin{aligned}
\exp\left(a\mathbb{E}_{\sigma\in\{\pm1\}^n}\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^n\sigma_i t_i\right)\right) &= \exp\left(\mathbb{E}_{\sigma\in\{\pm1\}^n}\left(a\sup_{t\in\mathcal{T}}\sum_{i=1}^n\sigma_i t_i\right)\right)\\
&\leq \mathbb{E}_{\sigma\in\{\pm1\}^n}\left(\exp\left(a\sup_{t\in\mathcal{T}}\sum_{i=1}^n\sigma_i t_i\right)\right)\\
&= \mathbb{E}_{\sigma\in\{\pm1\}^n}\left(\sup_{t\in\mathcal{T}}\left(\exp\left(a\sum_{i=1}^n\sigma_i t_i\right)\right)\right)\\
&\leq \sum_{t\in\mathcal{T}}\mathbb{E}_{\sigma\in\{\pm1\}^n}\left(\exp\left(a\sum_{i=1}^n\sigma_i u t_i\right)\right),
\end{aligned}$$

where for the first inequality we have used Jensen's inequality and the second equality holds due as $\exp(\cdot)$ is strictly monotonically increasing. The right-hand side is just an MGF which can be split into a product due to independence, hence

$$\begin{aligned}
\exp\left(a\mathbb{E}_{\sigma\in\{\pm1\}^n}\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^n\sigma_i t_i\right)\right) &= \sum_{t\in\mathcal{T}}\prod_{i=1}^n\mathbb{E}_{\sigma_i}\left(\exp\left(a\sigma_i t_i\right)\right)\\
&\leq \sum_{t\in\mathcal{T}}\prod_{i=1}^n\exp\left(\frac{a(2t_i)^2}{8}\right),
\end{aligned}$$

where we get the inequality from Lemma 2.1.3. Therefore,

$$\begin{aligned}
\exp\left(a\mathbb{E}_{\sigma\in\{\pm1\}^n}\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^n\sigma_i t_i\right)\right) &\leq \sum_{t\in\mathcal{T}}\exp\left(\frac{a^2}{2}\sum_{i=1}^n t_i^2\right)\\
&\leq \sum_{t\in\mathcal{T}}\exp\left(\frac{a^2\max_{t\in\mathcal{T}}\|t\|^2}{2}\right)\\
&= \exp\left(\frac{a^2\max_{t\in\mathcal{T}}\|t\|^2}{2}\right)\|t\|^2\left|\mathcal{T}\right|.
\end{aligned}$$

Taking the logarithm of both sides and dividing by a we get that

$$\mathbb{E}_{\sigma\in\{\pm1\}^n}\left(\sup_{t\in\mathcal{T}}\left(\sum_{i=1}^n\sigma_i t_i\right)\right) \leq \frac{\log\left(|\mathcal{T}|\right)}{a} + \frac{a\max_{t\in\mathcal{T}}\|t\|^2}{2} = \max_{t\in\mathcal{T}}\|t\|\sqrt{2\log\left(|\mathcal{T}|\right)},$$

which completes the proof of the lemma. ■

Proof. Let $T \subseteq \mathcal{G}$ be an ϵ-net of size $N(\mathcal{G},\rho,\epsilon)$, then by Lemma 6.11.1 we have that

$$\begin{aligned}
\mathbb{E}_{\sigma\in\{\pm1\}^m}\left(\max_{g'\in T}\frac{1}{m}\sigma_i g'(x_i,y_i)\right) &\leq \max_{g'\in T}\|g(x_i,y_i)\|_2\frac{\sqrt{2\log\left(N(\mathcal{G},\rho,\epsilon)\right)}}{m}\\
&\leq \sqrt{m}\frac{\sqrt{2\log\left(N(\mathcal{G},\rho,\epsilon)\right)}}{m}\\
&= \sqrt{\frac{2\log\left(N(\mathcal{G},\rho,\epsilon)\right)}{m}}.
\end{aligned}$$

Using this we can conclude that,

$$\begin{aligned}
\hat{\mathfrak{R}}(\mathcal{G}) &= \mathbb{E}_{\sigma\in\{\pm1\}^m}\left(\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(x_i,y_i)\right)\right)\\
&\leq \mathbb{E}_{\sigma\in\{\pm1\}^m}\left(\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(x_i,y_i)-\sigma_i g'(x_i,y_i)\right)\right)+\mathbb{E}_{\sigma\in\{\pm1\}^m}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g'(x_i,y_i)\right)\\
&\leq \mathbb{E}_{\sigma\in\{\pm1\}^m}\left(\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}|g(x_i,y_i)-g'(x_i,y_i)|\right)\right)+\mathbb{E}_{\sigma\in\{\pm1\}^m}\left(\max_{g'\in T}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g'(x_i,y_i)\right)\right)\\
&\leq \sup_{g\in\mathcal{G}}\rho((g(x_1,y_1),\ldots,g(x_m,y_m)),(g'(x_1,y_1),\ldots,g'(x_m,y_m)))+\sqrt{\frac{2\log\left(N(\mathcal{G},\rho,\epsilon)\right)}{m}}\\
&\leq \epsilon+\sqrt{\frac{2\log\left(N(\mathcal{G},\rho,\epsilon)\right)}{m}},
\end{aligned}$$

which holds for all $\epsilon>0$ which completes the proof of the theorem. □

References

[1] Michel Habib. “Probabilistic methods for algorithmic discrete mathematics”. In: 1998.

[2] David A. McAllester. “PAC-Bayesian model averaging”. In: *Annual Conference Computational Learning Theory*. 1999.

[3] John Langford and Matthias Seeger. “Bounds for Averaging Classifiers”. In: (Feb. 2001).

[4] Andreas Maurer. “A Note on the PAC Bayesian Theorem”. In: *CoRR* (2004).

[5] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[6] Gilles Blanchard and François Fleuret. “Occam’s Hammer”. In: *Learning Theory*. Ed. by Nader H. Bshouty and Claudio Gentile. Springer Berlin Heidelberg, 2007, pp. 112–126.

[7] Olivier Catoni. “Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning”. In: *IMS Lecture Notes Monograph Series* 56 (2007), pp. 1–163.

[8] Olivier Catoni. “A PAC-Bayesian approach to adaptive classification”. In: (Jan. 2009).

[9] Maria-Florina Balcan. *Rademacher Complexity*. 2011.

[10] David A. McAllester. “A PAC-Bayesian Tutorial with A Dropout Bound”. In: *CoRR* (2013).

[11] Clayton Scott. *Hoeffding’s Inequality*. 2014.

[12] Clayton Scott. *Rademacher Complexity*. 2014.

[13] Clayton Scott. *The Bounded Difference Inequality*. 2014.

[14] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. *Towards the Limit of Network Quantization*. 2017.

[15] 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).

[16] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. “A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks”. In: *CoRR* (2017).

[17] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. “Stronger generalization bounds for deep nets via a compression approach”. In: *CoRR* (2018).

[18] Benjamin Guedj. *A Primer on PAC-Bayesian Learning*. 2019.

[19] 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.

[20] Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, and Daniel M. Roy. “On the role of data in PAC-Bayes bounds”. In: *CoRR* (2020).

[21] Martin Lotz. *Covering Numbers*. 2020.

[22] Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvari, and John Shawe-Taylor. “PAC-Bayes Analysis Beyond the Usual Bounds”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 16833–16845.

[23] Pierre-Francois Rodriguez. *Lebesgue Measure and Integration*. 2021.

[24] Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. *A General Framework for the Disintegration of PAC-Bayesian Bounds*. 2021.

[25] Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. *PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization*. 2022.

[26] Patrick Rebeschini. *Algorithmic Foundations of Learning*. Nov. 2022.

[27] Pierre Alquier. *User-friendly introduction to PAC-Bayes bounds*. 2023.