USING OBJECTIVE BAYESIAN METHODS TO DETERMINE THE OPTIMAL DEGREE OF CURVATURE WITHIN THE LOSS LANDSCAPE

Anonymous authors
Paper under double-blind review

ABSTRACT

The efficacy of the width of the basin of attraction surrounding a minimum in parameter space as an indicator for the generalizability of a model parametrization is a point of contention surrounding the training of artificial neural networks, with the dominant view being that wider areas in the landscape reflect better generalizability by the trained model. In this work, however, we aim to show that this is only true for a noiseless system and in general the trend of the model towards wide areas in the landscape reflect the propensity of the model to overfit the training data. Utilizing the objective Bayesian (Jeffreys) prior we instead propose a different determinant of the optimal width within the parameter landscape determined solely by the curvature of the landscape. In doing so we utilize the decomposition of the landscape into the dimensions of principal curvature and find the first principal curvature dimension of the parameter space to be independent of noise within the training data.

1 INTRODUCTION

When training a neural network we aim to find a parametrization which minimizes the variance of the data around the model's conditional mean value. A statistic which is reflective of this variance is known as a loss function and can be seen as creating a landscape mapping a model parametrization to a corresponding loss value. Thus, higher points in the landscape reflect higher loss values and a worse model parametrization. The saliency of other features of the loss landscape on the model performance are relatively less clear and in some cases are points of contention within the field. One such point is whether the width of a basin in the landscape surrounding a local minimum (we will also refer to this as the width of the minimum) is reflective of the ability of a model parametrization at the minimum to generalize to unseen data. It is a common notion that the wider the minimum in the landscape, as measured by the Hessian matrix of the loss function (Keskar et al., 2016; Dinh et al., 2017), the better the model parametrization will generalize. The intuition behind such a belief is simply that, wider minima reflect that a model will experience less deviation in its loss metric as a result of minor deviations of its parameter values, and as a result the model is more robust than if it were to be parametrized by a very specific parameter set found at a sharp minimum.

In this work we aim to demonstrate that the width of a minimum is a key feature of the loss landscape and provides significant information on the progress of the training of a model. We deviate, however, from the views of the field that the widest minima provide the best generalizability by reflecting that there is instead an optimal width or curvature around the parametrization with the best generalizability which is not necessarily the widest point in the landscape. To this end Section 2 provides the necessary background information which we will utilize in developing our theories which are presented in Section 3. Section 4 and Section 5 then provide empirical evidence in support of the theoretical findings with Section 4 describing the methods employed to test the theories. Section 5 then provides and discusses the results of these empirical tests. Finally we conclude in Section 6 with our closing remarks.

The contributions of this work are threefold. Firstly we evaluate the concept of Energy-Entropy competition of neural networks (Zhang et al., 2018) in the context of the bias-variance trade-off (Geman et al., 1992) and reflect that a correlation exists between energy and entropy as opposed to a competition or trade-off as was first presented. Secondly we utilize the Fisher Information of the loss landscape in the area of a minimum to reflect that an optimal level of curvature exists within the landscape which does not necessarily occur at the point in the landscape with the least curvature. Further to this, we provide a novel view on the overfitting of models to their training data using the loss landscape. Finally, the Fisher Information is utilized in defining the objective Bayesian prior known as the Jeffreys prior and we show that at the point in the landscape which corresponds to the Jeffreys prior the test error of the model reaches its minimum value and at this point the dimension of principal curvature of the model is at its maximum entropy. In doing so we also reflect the noise invariance of the dimension of principal curvature.

2 Background

2.1 Fisher Information and Uninformative (Jeffreys) Priors

The Fisher Information (which we denote by $\zeta(\theta)$) is a metric dependent on the model parametrization which reflects the amount of information that a sufficient statistic based on the observable data, such as the variance of the data around the model predictions (Jaynes, 2003), carries about an unknown parameter θ. In the case of a Gaussian model, the Fisher Information is equal to the Hessian of the log-likelihood of the Gaussian. One of the key properties of the Fisher Information Matrix is that its determinant is invariant under reparametrizations of a trained model. Thus, when the parameters used in modelling a distribution are changed, the Fisher Information (and thus the Hessian in the Gaussian case) in each dimension will change, however, the determinant or volume of information remains unchanged between the model parametrizations.

The invariance property of the Fisher Information was the reason for its utilization in Jeffreys (1946) who sought to create a Bayesian prior with such an invariance property. The resultant prior is known as the Jeffreys prior and is shown in Equation 1.

$$P(\theta) \propto \sqrt{\det \zeta(\theta)} = \sqrt{\det H(\theta)} \tag{1}$$

As has been shown in Jaynes (1968; 2003) the utility of the Jeffreys prior is not limited to the invariance property, as the Jeffreys prior is an example of an uninformative or objective prior, and as a result informs the posterior distribution as little as possible. The Jeffreys prior is thus used to reflect complete prior ignorance about the correct model parametrization, resulting in a posterior distribution with parameters completely determined by the observed data.

A key perspective on the Jeffreys prior presented in Jaynes (2003) is that the prior imposes a uniform distribution over the function space of the model, not the parameter space. This distinction is made clearer by an example from Jaynes (2003). Assume we are modelling a function of a million parameters $f(\theta_1, \theta_2, ...)$ and this function can only take a given value x with a single combination of the parameter values. Assume now that we observe this value x in our training data. We may simultaneously determine all one million parameters with this one data point. More importantly any prior which places full determination of the model parameters on the data would tell us to trust such a parametrization as this would be the only means of seeing such a data point. The inverse is true if we observe model behaviour which could be the result of a number of parametrizations. It would be necessary to reflect less certainty on any of the parametrizations which cause such behaviour, merely because of the many other indistinguishable parametrizations given the data. We see that the prior places higher density on parametrizations with a unique function approximation and less density on parametrizations which provide a function approximation which could be similarly achieved by another parametrization. This would result in a uniform prior over the function approximations and as a result the choice between function approximations is left to the model learning from the data.

2.2 The Bias-Variance Dilemma, Energy-Entropy Competition and Minimum Description Length

It is a well-known fact that a model learning to equate its conditional mean precisely to the values found in the training data is not always beneficial to the performance of the model on unseen data. In particular when we observe a decrease in training error but increase in validation or test error we say that the model is overfitting the training data (Hawkins, 2004). In Geman et al. (1992) it is shown empirically that to decrease the variance of the data around a model's predictions (reduce the training error) it is necessary for the variance in the model parameters to increase. Further, Geman et al. (1992) reflect that a large parameter variance corresponds to the overfitting of the model to the training data. This trade-off between the bias of the model and the variance of its parameters is known as the Bias-Variance Dilemma (Sammut & Webb, 2011). In Geman et al. (1992) the only means presented to mitigate this dilemma is to obtain more training data.

We see, however, that neural networks are capable of learning complex tasks with limited data and even generalize in spite of the Bias-Variance Dilemma. In Zhang et al. (2018) it is argued that the success of neural networks is due to a bias of Gradient Descent towards wider minima in the loss landscape. To reflect this, Zhang et al. (2018) derive a Bayesian posterior distribution for the parameters of a model given the training data. To derive this distribution, Zhang et al. (2018) utilize a Gaussian likelihood with conjugate Gaussian prior, which we generalize in Equation 2 by allowing any prior distribution $exp(h(\theta))$ which results in a proper posterior distribution. The exponential term of this generalized prior $h(\theta)$ is seen as some function of θ while $f(x_i, \theta)$ denotes the function approximation by the neural network, σ_i^2 is the variance of the output corresponding to data point x_i, y_i is the true output for a particular x_i in the training data and finally $H(\theta)$ denotes the Hessian matrix for parametrization θ. The derivation of Equation 2 can be found in Appendix A.1.

$$P(\theta|\mathbf{X}) = \frac{1}{Z} \exp\left[-\left(\sum_{i=1}^{P} \frac{(y_i - f(x_i, \theta))^2}{2\sigma_i^2} - h(\theta) + \frac{1}{2} \log det(H(\theta)) \right) \right] \tag{2}$$

From Equation 2 we see that to maximize the probability of a parametrization we must maximize the model likelihood (minimize the first term in the exponential), maximize the prior probability of the parametrization and maximize the model entropy which is reflected by the final term in the exponential and is inversely proportional to the Hessian of the loss landscape at the parametrization. Using their conjugate Gaussian posterior distribution Zhang et al. (2018) note that maximizing the model likelihood is not the only factor which should be used in determining the model parametrization, and in some cases it may be beneficial to trade-off some training error for an increase in model entropy, which the authors called Energy-Entropy competition. Zhang et al. (2018) state that the bias of Gradient Descent towards wider minima, with smaller Hessian values, results in the model naturally maximizing entropy, aiding in its generalizability. We see, however, from the Bias-Variance Dilemma that by reducing the bias of the model on the training data, and thus increasing its likelihood, that the model entropy will naturally also increase due to the higher variance in the parameter values.

With the perspective of both the Bias-Variance Dilemma and Energy-Entropy competition we see that wide points in the loss landscape have been related to both overfitting and improved generalizability of a model parametrization. Thus, from one perspective we aim for sharp minima within the landscape and from the other we should aim for wide minima. The issue of the width of a minimum is further confounded by Dinh et al. (2017) which states that the width of a minimum is not a consistent indicator of the ability of a model to generalize. The impact of the width of a minimum in the landscape is, still an open question and one which we try address in this work.

The Minimum Description Length (MDL) Principle is an information theoretic principle which states that the optimal model for a set of data provides the best compression of the data (Rissanen, 1978). In other words the optimal model is the simplest model which incurs the least training error. This principle is another example of the trade-off between model complexity and minimizing the model bias. Due to its assertion that the optimal model is the simplest unbiased one the MDL Principle is the mathematical formulation of Occam's Razor, and is expressed by Equation 3, which reflects that to compress the data D optimally we must find the parametrization with the net minimum entropy in the parameter space $L(\theta)$ and in the data space given the parametrization $L(D|\theta)$.

$$L(D) = \min_{\theta \in \vartheta} (L(\theta) + L(D|\theta)) \tag{3}$$

For the exponential family of likelihood distributions the Jeffreys prior is used to enforce the MDL property on the posterior distribution and results in a minimax optimal posterior, which is to say that the maximal risk of the model parametrization is minimal out of all unbiased parametrizations (Lehmann & Casella, 2006). The minimax optimality property, thus, provides the lowest upper bound of the risk for all model parametrizations. Thus the MDL property is related to the Bias-Variance Dilemma and MDL posterior distributions aim to avoid overfitting.

A final necessary principle which encompasses all of the topics above is the Likelihood Principle (Jaynes, 2003), which states that within the context of a specified model, the likelihood distribution $L(D|\theta)$ trained from data D contains all the information about the model parameters θ that is contained in D.

2.3 PRINCIPAL CURVATURE

The Jeffreys prior, and by extension the Fisher Information, finds further utility in its use as a right Haar measure for the parameter space of a normal distribution (Berger, 2013). The Haar measure is used to assign an invariant volume to a subset of locally compact topological groups and, thus, form an integral over the compact groups. In the case of the parameter space of the normal distribution the topological groups will be of parametrizations with similar function approximations and, thus, similar loss metric within the basin surrounding a local minimum in the loss landscape. Further, we note that the parameter space of a probabilistic model forms a statistical manifold and by extension a Riemannian manifold (Rao, 1945). The metric tensor for statistical manifolds is the Fisher Information metric (Skovgaard, 1984), defined as the expected value of the individual elements of the Fisher Information matrix, which forms the tangent space of such manifolds. As stated in Section 2.1, in the case of Gaussian parameter spaces the Fisher Information Matrix can be equally derived as the Hessian matrix of the loss function relative to the model parameters. This is significant as the Hessian matrix is used in the area of a critical point on a Riemannian manifold for obtaining the shape operator (Spivak, 1970), and as a result the principal curvatures at the point (Porteous, 2001). In the case of a Gaussian parameter space the shape operator is the Gaussian curvature defined as the determinant of the Hessian matrix $det(H(\theta))$ (Koenderink & Van Doorn, 1992). The principal curvatures are defined as the eigenvalues of the Hessian matrix and decompose the manifold into orthogonal dimensions of curvature, with the first eigenvector reflecting the dimension of most curvature.

It is important to note that while the parameter space of a statistical model forms a Riemannian manifold, when parametrized by an overly-determined model such as a neural network, the parameter space will not be Riemannian but rather semi-Riemannian, due to the fact that the Fisher Information metric will no longer be defined over the entire manifold. Such undefined points for the metric are a result of a singular metric tensor at the model parametrization and occur due to the covariance of parameters within the model. Covariant parameters necessarily occur with the addition of hidden layers to the model and result in dimensions on the manifold in which the parameters may be varied without altering the behaviour of the model. This results in dimensions of no curvature along the manifold. As seen in Section 5, this is not a destructive point for the training procedure, however, we must necessarily remain cognisant of such covariant dimensions along the statistical manifold.

3 MODEL ENTROPY, THE LOSS LANDSCAPE AND GENERALIZATION

The aim of training a neural network is to find the most probable parametrization for a model as determined by the posterior probability reflected by Equation 2. This is achieved by maximizing the combination of the likelihood, prior probability and entropy of the model parametrization. The likelihood we increase normally by decreasing the variance of the training data around the model predictions. The entropy term we have no direct control over as the landscape is completely determined by the data, the sufficient statistic being used to determine the parameters (which is the loss metric) and the hypothesis (the model architecture being trained). So the only component of the posterior left to be determined is the prior. As with most work in Bayesian statistics this is the most difficult part and must be treated with great care. There are presently two common approaches to setting this prior distribution, the first of which being to not specify one, or more precisely use an implicit uniform prior (Chaudhari & Soatto, 2018) and, thus, use maximum likelihood estimation to determine the parameter values. The second common approach is to utilize a conjugate Gaussian with a mean of 0 for the prior. In practice this method takes the form of L2 regularization, also known as weight decay (Krogh & Hertz, 1992), with $h(\theta) \propto ||\theta||^2$ in Equation 2, the approach employed by Zhang et al. (2018). Neither approach has proven to be sufficient consistently for deterring models from overfitting, without introducing a form of bias, due to their unjustified assumptions about the correct model parametrization. This is relatively clear in the conjugate Gaussian prior approach which assumes a mean of 0 for the parameter values. In a case where we have explicit prior knowledge that such a mean and distribution is in-fact correct for the model parameters then this would be a correct approach, however, in almost every case we are completely ignorant to the values of the true parametrization and, thus, we would be biasing our models to some degree by using this prior. It is necessary to developing an unbiased model that this absence of prior knowledge be reflected in the training procedure.

The source of error from the uniform prior is slightly more nuanced, as initial intuition would suggest that giving equal probability to all values a parameter could take is a correct means of reflecting our prior ignorance about the parameter values. However, this method fails to accurately reflect the probabilities that unlikely parametrizations may be correct and in a sense exhibits a kind of confirmation bias. To state this another way, the uniform prior over the parameter space is not the uniform prior over the function space. This is reasonably what would be required, however, as if two model parametrizations result in equivalent behaviour this should reduce our certainty when determining one or the other parametrization as being the true parametrization, after observing said behaviour in our data. Likewise if we know that only one possible parametrization can exhibit a certain behaviour and we observe the given behaviour we can be relatively more confident in the parametrization we determine to be true. By placing equal prior probability on all areas of the landscape we see that the use of the uniform prior will result in a posterior distribution over the model parameters which places excessive density on high variance areas of the landscape while at the same time places too little probability on very specific, low variance parametrizations of the model, resulting in a model which favours wider minima within the loss landscape.

How then would one go about determining what prior probability to assign to a parametrization without knowledge of the other possible parametrizations which would be equally as likely? We would be inclined to use a form of variance in the parameter space, as a high variance in parameter space would reflect that given the training data, sufficient statistic and hypothesis there were many possible parametrizations which could be found with similar likelihood values. We note that this is exactly what the entropy term in Equation 2 is telling us. Higher entropy means wider minima which reflect higher parameter variance and thus the necessity to be less certain of the parametrization in that area. The opposite is true for an increase in certainty in our parametrization at a sharp minimum. It is thus possible to utilize the loss landscape in the area of a minimum to determine the degree of certainty we may have in our model parametrization being the true parametrization and as a result determine the necessary prior probability. Furthermore, this would mean we are setting our prior based on the model behaviour given the observed data and sufficient statistic. Note, we do not say we determine our prior based on the hypothesis as the determinant of the Fisher Information/Hessian is invariant under reparametrizations. This means that in the area of a minimum, by transforming the hypothesis to be modelled by an alternate set of parameters θ' the dimensions and volume of the landscape will adjust such that $\sqrt{det(H(\theta))} = \sqrt{det(H(\theta'))}$ (Fisher, 1922). In conclusion we may set the prior probability to be equal to or proportional to the entropy determined by the loss landscape. Thus we conjecture that a correct prior for a model would be:

$$P(\theta) \propto \sqrt{\det \zeta(\theta)} = \sqrt{\det H(\theta)} \tag{4}$$

As a result we see in Equation 2 such a prior would give the equation for $h(\theta)$ as

$$h(\theta) = \frac{1}{2} \sum_{i=1}^{N} \log det(H(\theta)) \tag{5}$$

In such a case the prior and entropy term in the posterior formulation cancel out, leaving the likelihood term as the only factor determining the posterior probability, as can be see in Equation 6.

$$P(\theta|\mathbf{X}) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{P} \frac{(y_i - f(x_i, \theta))^2}{2\sigma_i^2}\right) \tag{6}$$

Significantly we also see that Equation 4 is the Jeffreys prior as shown in Equation 1. The fact that the likelihood is the only term which remains in the posterior distribution appears to be in agreement with the notion that the Jeffreys prior reflects complete prior ignorance and places all determination of the parameters on the data at hand.

A necessary distinction regarding the Jeffreys prior is that, while it places the full parameter determination on the data, it does not necessarily result in a posterior distribution which has extracted all information from the data. Any information which provides an insufficient decrease in data variance to warrant the increase in variance in the model parameters will not be utilized as the model naturally "distrusts" this information by providing a relatively lower prior probability to the more entropic parametrization found in the wider basin. This is where we see the utility of the Jeffreys prior becoming evident with regard to the MDL property reflected by equation 3. As discussed in Section 2.2 the model which is optimal is the one which has the lowest variance of the data around its predictions $L(D|\theta)$ while maintaining low variance in the model parameter space $L(\theta)$. As also discussed in Section 2.2 the bias-variance trade-off (Geman et al., 1992) reflects that to decrease data variance we must increase parameter variance reflecting the trade-off between $L(\theta)$ and $L(D|\theta)$ in the MDL equation. Further, we related this to the Energy-Entropy competition concept (Zhang et al., 2018), where it was also stated that neural networks are biased towards wide minima. The wide minima bias is exactly in agreement with our discussion above whereby a uniform prior, and thus, maximum likelihood estimation is biased toward finding the widest possible local minimum. In the notation of the MDL equation we see that maximum likelihood estimation aims only to minimize $L(D|\theta)$, regardless of $L(\theta)$. In this case, no such distrust of high variance parametrizations is found. We see, however, in the information theory literature that the Jeffreys Prior is used to enforce the MDL condition for the exponential family of distributions (Rissanen, 1978), which supports our assertion that the Jeffreys prior is the correct prior to use to reflect the absence of prior knowledge about the model parameters as it balances both model complexity $L(\theta)$ while fitting the data $L(D|\theta)$.

The primary power of the Jeffreys prior comes from the use of the Fisher Information. Naturally as the model fits the data and captures information, the amount of information left in the data which remains uncaptured by the model decreases. This is observed as a decrease in the Fisher Information. We see, however, that the model entropy increases as the Hessian matrix, and by extension the Fisher Information, decreases, which is again in agreement with the bias-variance trade-off. The consequence of this observation is that to capture all the information from a sufficient statistic determined by the training data we must utilize increasingly complex models, capable of modelling finer details found in the data. The utility of such fine details to the model performance exhibits diminishing returns to a point where the perturbation of a parameter capturing these details will not result in any significant deviation in the model behaviour. Simply, as more information is shared between parameters, the individual importance of a parameter decreases. This is in contrast to an under-parametrized model where the parameters capture as much of the most important information from the sufficient statistic as possible and rely heavily on this information in determining its behaviour. In light of the Fisher Information we see the Maximum Likelihood Principle further reflects that the use of the uniform prior biases our models towards maximum entropy within the loss landscape by extracting all information from the training data at the expense of higher model complexity.

It must be noted that the propensity of neural networks to extract all information from the training data is not an inherently negative quality of the models. Quite the opposite, it is reflective of the power and capability of the models which are designed to learn the variances within data and utilize this information in determining their behaviour. As a result, however, the efficacy of these models is directly related to the reliability of the data on which they are trained and for all the information found in the training data to be present and reflective of the entire population of data being modelled. This is seldom the case as training data is inevitably noisy, either due to noise from sampling and capturing of the data, or due to confounding aspects of the task domain which on average do not affect the population data distribution but do provide a source of structured noise when their effect is observed on the training data. Minimizing the Fisher Information metric and fulfilling the Maximum Likelihood Principle in such cases would reflect that the information found in the noise of the training data was utilized in determining the model parameter values, which is clearly undesirable and is known as the model over-fitting the training data (Hawkins, 2004).

We thus see that the bias of maximum likelihood estimation to the widest areas in the loss landscape is indicative of the propensity for maximum likelihood models (including unregularized neural networks) to overfit the data on which they are trained. Further, we see that the notion of the widest possible minima in the loss landscape providing the best generalization performance is only true in a noiseless environment. The view, however, that wide areas in the landscape generalize better is true as this width in the landscape would merely reflect that the model has captured more information from the data than a parametrization found in a steeper portion of the landscape. Naturally this would provide better out-of-sample performance by the model if it has captured the information found in the training data which is reflective of the information within the population data distribution which we aim to model. We conclude that the width of the landscape in which a model finds itself is demonstrably important and that there is a precise width in the landscape which provides the model with the best possible out-of-sample performance. This point would be exactly where the model prior is equal to the Jeffreys Prior as determined by the Fisher Information of the loss landscape. This is due to the aversion of the Jeffreys prior to any information which does not justify the increase in model entropy by a superior decrease in data variance or prediction error, while remaining objective in the sense that the prior is completely determined by the loss landscape and has as little effect on the parameter posterior distribution as possible. Thus, noise in the training data which only serves to perturb and hinder the learning of the model without providing sufficient benefit to how the training data is fit will not be learnt by the model.

4 METHODS

To empirically confirm the theoretical findings presented in Section 3 it was necessary to have knowledge of the true underlying distribution or physical “model” which was generating the data. To achieve this perspective a data generating model was used. This model was a neural network composed of one, 5 neuron, hidden layer which utilized a sigmoid activation function in its hidden layer and a linear activation in its scalar output layer. The true parameters were independently sampled from a Gaussian distribution, centered around 0 with covariance 0.8, for each training procedure. The sigmoid activation in the hidden layer ensured that a complex non-linear distribution was constructed. To generate the data on which another model would be trained we generated all 25-bit strings and input this data into the generating model, capturing the corresponding scalar outputs.

As stated in Section 3, however, if the model were to be trained on this data the maximum entropy bias would be of assistance to the learning procedure as it would ensure the model extracts all possible information from the data, and in such a case all information would be accurate and reliable. Thus, it was necessary to add Gaussian noise to each output of the original regression. The noise was sampled with a mean of 0 and covariance ranging between 0.2 and 0.4 for different trainings. It is important to note the scale of the noise was relatively small compared to the scale of the actual output values. As discussed in Section 2.2 the MDL principle states that an optimal model balances complexity with minimizing training bias. Naturally modelling such noise would require a higher degree of model complexity, however, if too much noise was added then it would be beneficial for a model to accept the higher model complexity to fit the noise and reduce its bias. In such a case it would be as if the training data was sampled from a separate population of data to the test data, and observing the training dynamics of this experiment would not aid in understanding overfitting. In total $33\,554\,432$ input-output data pairs were generated from the noisy true model function of which 50 were used to train the model. The small training sample encouraging the model to overfit. The remainder of the data was used to measure the out-of-sample or test performance of the model.

The above generated dataset was then used to train a larger model of 5 layers with the first two layers utilizing a sigmoid activation and the remaining layers using linear mappings. The input and hidden layers contained 25, 20, 12 and 7 neurons respectively, with the output layer naturally having a single neuron for the regression task. While training the model the training error and test error were calculated after each update step, with the aim of detecting the precise location where the test error began to increase while the training error continued decreasing. It was also necessary to utilize a metric for the difference between the model and true distribution and the probabilities of the respective parametrizations. For this task we utilized the Jeffreys divergence, show in Equation 7 below. The Jeffreys Divergence is merely the sum of the KL-divergence for the true parameter distribution $T(\theta)$ compared to the model parameter distribution $P(\theta|\mathbf{X})$ and the opposite KL-divergence. We show the derivation of the Jeffreys Divergence from this summation in Section A.2 of the Appendix.

$$D_J(T(\theta)||P(\theta|\mathbf{X})) = \int (T(\theta) - P(\theta|\mathbf{X}))(lnT(\theta) - lnP(\theta|\mathbf{X}))d\theta \tag{7}$$

Naturally Equation 7 is intractable due to the necessity to integrate over all parametrizations. However, as discussed in Section 3, the use of maximum likelihood estimation results in an excess of density on high variance parametrizations in the posterior parameter distribution in Equation 2. It was thus sufficient to evaluate the Jeffreys Divergence at a single point in parameter space and observe the relative densities at that point, providing a distance metric as opposed to a divergence metric. We will, thus, refer to the Jeffreys Distance for the remainder of this work, with the formula shown in Equation 8.

$$D_J(T(\theta)||P(\theta|\mathbf{X})) = (T(\theta) - P(\theta|\mathbf{X}))(lnT(\theta) - lnP(\theta|\mathbf{X})) \tag{8}$$

The necessity of a distance metric being positive semi-definite is upheld by this metric as it is clear when $(T(\theta) - P(\theta|\mathbf{X})) < 0$ then $(lnT(\theta) - lnP(\theta|\mathbf{X})) < 0$. Likewise when $(T(\theta) - P(\theta|\mathbf{X})) > 0$ then $(lnT(\theta) - lnP(\theta|\mathbf{X})) > 0$. This is a benefit of the symmetrical property of the Jeffreys Divergence which the KL-divergence does not possess. Thus, substituting the model posterior formula from Equation 2 as well as the true model distribution in Equation 9 into the logarithmic term of Equation 8 we obtain the formulation shown in Equation 10.

$$T(\theta^*) = \frac{1}{Z*} \exp\left(-\sum_{i=1}^{P} \frac{(y_i - f(x_i, \theta^*))^2}{2\sigma_i^2}\right) \tag{9}$$

$$(lnT(\theta) - lnP(\theta|\mathbf{X})) = \left(-\sum_{i=1}^{P} \frac{(y_i - f(x_i, \theta^*))^2}{2\sigma_i^2} + \sum_{i=1}^{P} \frac{(y_i - f(x_i, \theta))^2}{2\sigma_i^2} - h(\theta) + \frac{1}{2}\log det(H(\theta)) + Z^* - Z\right) \tag{10}$$

Assuming now that $\theta = \theta^*$, and thus $f(x_i, \theta) = f(x_i, \theta^*)$, as would be the case at the end of an unbiased training procedure, we see that the two variance terms in Equation 10 will cancel out. Further, we see that the only means of obtaining a 0 value for the expression is to use the Jeffreys prior causing $h(\theta)$ to cancel with the entropy term $\frac{1}{2}\log det(H(\theta))$, as discussed above in Section 3. Finally we see that using the Jeffreys prior would result in the posterior model distribution shown in Equation 6. If $f(x_i, \theta) = f(x_i, \theta^*)$ in this case then it is clear that $Z = Z^*$ is the necessary corresponding normalizing constant, taking care of the two normalizing constants in Equation 10. A similar argument can be made for the probabilities component of the Jeffreys distance $(T(\theta) - P(\theta|\mathbf{X}))$ in Equation 8, whereby we equate the two likelihood terms, then the use of the Jeffreys prior ensures that the posterior distributions do not differ resulting in a Jeffreys Distance of 0.

From Equation 10 we see that the use of the Jeffreys prior while minimizing the error of the model directs the model to a parametrization which results in the model likelihood being equal to that of the true underlying distribution. This allowed us to determine the point at which the model past the most probable parametrization in the landscape corresponding to the use of a Jeffreys prior, as being the parametrization which equated the two likelihood values. The first empirical result presented in Section 5 is thus a distribution reflecting the proximity of the training step at which minimum test error was reach to the training step at which the likelihood values intersected. A natural indication of a model overfitting would be for the model to become a more likely distribution to have generated the data than the true distribution itself. This is further highlighted by the Bias-Variance Dilemma which would state that for the model distribution to achieve a lower training error than the true distribution it must express a higher complexity, and model entropy, than the true model. Having observed the relationship between test error and the relative likelihoods, we then utilize the decomposition of the parameter landscape into its dimensions of principal curvature to reflect the impact of noise in the training data on the landscape by first training on noiseless data until a certain training error (error < 0.4) is achieved, at which point the noise is added to the data. We observe and interpret the resulting impact of the addition of the noise on the curvature of the landscape. Secondly we aim to observe the relationship between the dimensions of principal curvature and the generalizability of a model parametrization by observing the entropy on a Riemannian sub-manifold of the original statistical manifold (Chen, 2019), defined over the primary dimensions of principal curvature, relative to the test error and likelihood values of a model parametrization.

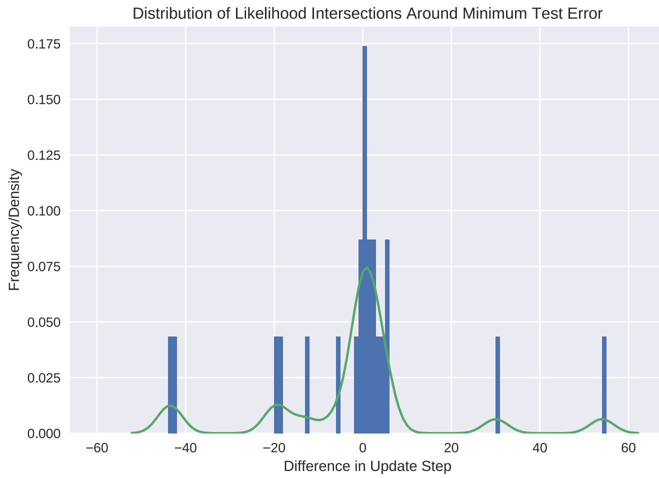

Figure 1: Distribution of the Intersection of the Likelihood Values in Equation 10 around the Minimum of the Test Error (n=23)

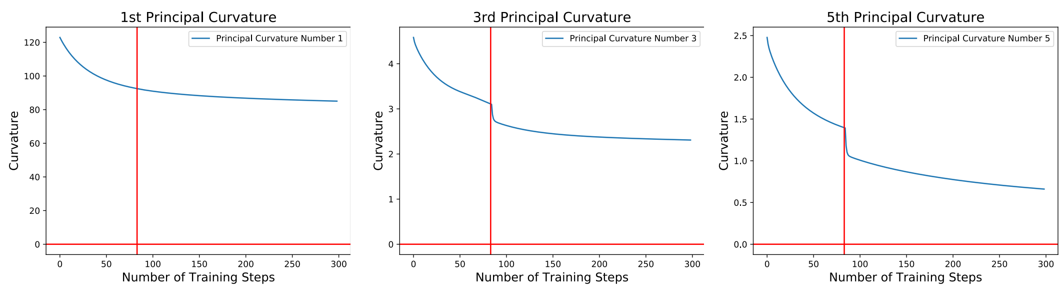

Figure 2: Impact of the addition of noise on 3 of the 5 Principal Curvatures of the Loss Landscape

5 RESULTS

As stated in Section 4 the first result aimed to determine if the point in the landscape which is most probable under the posterior distribution resulting from the use of the Jeffreys prior possess the optimal out-of-sample or test performance. The results of this first test are presented in Figure 1. We see in Figure 1 that the highest density is placed around 0 which supports the assertion that the Jeffreys prior results in the parametrization with the best generalization performance which is not necessarily in the widest area of the landscape. We do, however, see the appearance of some points outside of the cluster around 0. These points occur either due to the model learning noise from the data early in the training procedure, or due to imprecision in the learning rate used during training. Naturally if too large a learning rate is used we see the test error reach its minimum prior to the likelihood values intersecting, merely due to the inability of the model to extract the final bits of true information from the data within learning noise. It is also possible for the model to learn noise in the data before learning true information. If this occurs we will see the test error being minimized only after the likelihoods intersect as by learning the noise the model will increase its likelihood value on the training data, while the test error is only minimize (at a higher value than if no noise was captured) once sufficient true information had been captured. This is easily detectable in the training procedure as we can observe the test error stagnates or increase for the period when noise is learned, and then begin decreasing again.

The results of injecting noise into the training data only once the model has been sufficiently trained on clean data can be see in Figure 2. In this figure we reflect the Eigenvalues of 3 of the 5 principal curvatures of the loss landscape. Thus, each value reflects the inverse of the variance of the model in the dimension of the corresponding Eigenvector. Thus a lower value reflects a higher entropy in the given dimension. From these results we can see that the injection of noise results in a sudden

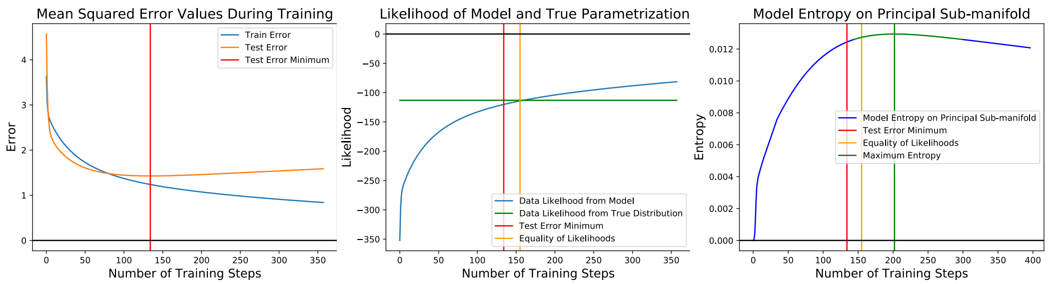

Figure 3: The Jeffreys prior parametrization is found in the area of parameter space minimizing test error and maximizing entropy on the principal sub-manifold

increase in the entropy of the lower principal curvature dimensions but has no effect on the first dimension of principal curvature. As the Fisher Information is reflected by the entropy this would mean that the low principal curvature dimensions capture more information at the injection of the noise into the data while the information captured by the first principal curvature dimension remains smooth throughout the training procedure. This would reflect a noise invariance of the principal curvature of the landscape and as a result we can see that the dimension of principal curvature in the landscape is exclusively responsible for the capturing of the true or primary data information. In light of this noise invariance we observed the entropy of a sub-manifold corresponding only to the dimensions of high curvature (Eigenvalues > 1) during the training of a model. The results of this test are reflected in Figure 3. In agreement with Figure 1 we see that in the area of the Jeffreys prior parametrization (equality in likelihood values), the test error reaches its minimum value. A number of insights can be gained from the third graphic in Figure 3. Firstly, that the Fisher Information matrix and the Fisher Information Metric, are non-singular and positive semi-definite on this sub-manifold. This reflects that the dimensions responsible for capturing true information in the data are convex, with positive Gaussian curvature and that it is sufficient for the model to merely minimize this well-behaved region of parameter space. We, further, observe that the Jeffreys prior parametrization maximizes the entropy of this sub-manifold, reflecting that it still captures all true information from the data. We see the green portion of the entropy metric as being the area where the entropy is within 0.003 of its maximum value. The fact that the entropy begins decreasing later in the training is reflective of the model forgetting true information while learning the noise once it starts overfitting. The fact that the entropy stagnates past the Jeffreys prior parametrization is due to the fact that even some of the principal curvatures were sensitive to noise and that in this region the model is beginning to learn noise and forget true information at the same rate. We, thus, see the final point that in the absence of noise capturing all information, maximizing entropy, is beneficial, however, in the presence of noise maximum entropy reflects overfitting.

6 Conclusion

We see that the notion of the width of the landscape being an indicator of a robust parametrization is correct, however, this is conditional upon the model being developed in a noiseless domain or, more significantly along a dimension of parameter space which is independent to the noise of the domain. With the aid of the Fisher Information perspective of the geometry of the landscape we see that the higher entropic points in the landscape directly reflect the absence of further information upon which the parameter values may be determined. Thus, we see the propensity of maximum likelihood models towards such high entropic points as a reflection of their propensity to utilize all information in determining the parameter values, including noise. Thus, we make the final conclusion that the point of maximum entropy in the loss landscape does not possess the best generalization performance and corresponds to the overfitting of the model to the training data. Instead the optimal point in the landscape occurs at maximum entropy in the dimension of principal curvature which corresponds to the most probable parametrization found by a Bayesian posterior distribution resulting from the use of the Jeffreys prior.

REFERENCES

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Bang-yen Chen. *Geometry of submanifolds*. Courier Dover Publications, 2019.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1019–1028. JMLR. org, 2017.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

Edwin T Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3): 227–241, 1968.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Jan J Koenderink and Andrea J Van Doorn. Surface shape and curvature scales. *Image and vision computing*, 10(8):557–564, 1992.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In J. E. Moody, S. J. Hanson, and R. P. Lippmann (eds.), *Advances in Neural Information Processing Systems 4*, pp. 950–957. Morgan-Kaufmann, 1992. URL `http://papers.nips.cc/paper/563-a-simple-weight-decay-can-improve-generalization.pdf`.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Ian R Porteous. *Geometric differentiation: for the intelligence of curves and surfaces*. Cambridge University Press, 2001.

C.R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37:81–91, 1945.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

Lene Theil Skovgaard. A riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, pp. 211–223, 1984.

Michael D Spivak. *A comprehensive introduction to differential geometry*. Publish or perish, 1970.

Yao Zhang, Andrew M Saxe, Madhu S Advani, and Alpha A Lee. Energy–entropy competition and the effectiveness of stochastic gradient descent in machine learning. *Molecular Physics*, 116 (21-22):3214–3223, 2018.

A Appendix

A.1 Derivation of the Neural Network Bayesian Posterior Distribution

The derivation presented below is adapted from Zhang et al. (2018).

We aim to derive a Bayesian posterior probability for the parameters of the model conditional upon the training data. We thus aim to obtain:

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{\sum_j P(\mathbf{X}|\theta)P(\theta)} = \frac{1}{Z}P(\mathbf{X}|\theta)P(\theta)$$

Assuming the data points are independently sample, we may assume a likelihood of:

$$P(\mathbf{X}|\theta) \propto \exp\left(-\sum_{i=1}^{P}\frac{(y_i - f(x_i,\theta))^2}{2\sigma_i^2}\right)$$

Apply a Gaussian Prior:

$$P(\theta) = \exp^{h(\theta)}$$

We obtain a posterior distribution:

$$P(\theta|\mathbf{X}) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{P}\frac{(y_i - f(x_i,\theta))^2}{2\sigma_i^2} + h(\theta)\right)$$

Utilizing Laplace Approximation for the integral of this posterior we obtain:

$$\int P(\theta|\mathbf{X})d\theta \approx \frac{1}{Z}\sum_q \frac{\exp\left[-\sum_{i=1}^{P}\frac{(y_i - f(x_i,\theta_\mathbf{q}))^2}{2\sigma_i^2} - h(\theta)\right]}{\sqrt{\det H(\theta_q)}}$$

where $\theta_\mathbf{q}$ are the local minima of a given loss function and $H(\theta_\mathbf{q})$ denotes the Hessian of the parameter matrix.

Hence, we write the denominator in exponential form:

$$\frac{1}{\sqrt{\det H(\theta_q)}} = \exp\left(-\frac{1}{2}\log det(H(\theta_\mathbf{q}))\right)$$

We obtain:

$$\int P(\theta|\mathbf{X})d\theta \approx \frac{1}{Z}\sum_q \exp\left[-\left(\sum_{i=1}^{P}\frac{(y_i - f(x_i,\theta_\mathbf{q}))^2}{2\sigma_i^2} - h(\theta_\mathbf{q}) + \frac{1}{2}\log det(H(\theta_\mathbf{q}))\right)\right]$$

With the point density for the model parametrization at a single minimum:

$$P(\theta|\mathbf{X}) \approx \frac{1}{Z}\exp\left[-\left(\sum_{i=1}^{P}\frac{(y_i - f(x_i,\theta_\mathbf{q}))^2}{2\sigma_i^2} - h(\theta_\mathbf{q}) + \frac{1}{2}\log det(H(\theta_\mathbf{q}))\right)\right]$$

A.2 Derivation of Jeffreys Divergence

$$
\begin{aligned}
D_J(T(\theta)||P(\theta|\mathbf{X})) =& D_{KL}(T(\theta)||P(\theta|\mathbf{X})) + D_{KL}((P(\theta|\mathbf{X})||T(\theta)) \\
=& -\int T(\theta) ln\left[\frac{P(\theta|\mathbf{X})}{T(\theta)}\right] d\theta - \int P(\theta|\mathbf{X}) ln\left[\frac{T(\theta)}{P(\theta|\mathbf{X})}\right] d\theta \\
=& -\int T(\theta)\left[lnP(\theta|\mathbf{X}) - lnT(\theta)\right] d\theta - \int P(\theta|\mathbf{X})\left[lnT(\theta) - lnP(\theta|\mathbf{X})\right] d\theta \\
=& \int -T(\theta)ln(P(\theta|\mathbf{X})) + T(\theta)ln(T(\theta)) - P(\theta|\mathbf{X})ln(T(\theta)) + P(\theta|\mathbf{X})ln(P(\theta|\mathbf{X}))d\theta \\
=& \int (-T(\theta) + P(\theta|\mathbf{X}))ln(P(\theta|\mathbf{X})) - (-T(\theta) + P(\theta|\mathbf{X}))ln(T(\theta))d\theta \\
=& \int (-T(\theta) + P(\theta|\mathbf{X}))(lnP(\theta|\mathbf{X}) - lnT(\theta))d\theta \\
=& \int (T(\theta) - P(\theta|\mathbf{X}))(lnT(\theta) - lnP(\theta|\mathbf{X}))d\theta
\end{aligned}
$$