Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 5843–5848

On the validity of using the delta method for calculating the uncertainty of the predictions from an overparameterized model ⋆

Magnus Malmström *, **Isaac Skog** **, **Daniel Axehill** *, **Fredrik Gustafsson** *

* *Linköping University, Sweden , (firstname.lastname@liu.se)*

** *Uppsala University, Sweden, (firstname.lastname@angstrom.uu.se)*

Abstract: The uncertainty in the prediction calculated using the delta method for an overparameterized (parametric) black-box model is shown to be larger or equal to the uncertainty in the prediction of a canonical (minimal) model. Equality holds if the additional parameters of the overparameterized model do not add flexibility to the model. As a conclusion, for an overparameterized black-box model, the calculated uncertainty in the prediction by the delta method is not underestimated. The results are shown analytically and are validated in a simulation experiment where the relationship between the normalized traction force and the wheel slip of a car is modelled using e.g., a neural network.

Keywords: Machine learning, nonlinear system identification, overparameterized model, uncertainty quantification, neural networks, autonomous vehicles

1. INTRODUCTION

This paper investigates how overparameterization affects the, via the delta method, calculated uncertainty in the prediction from a parametric black-box model, such as a neural network (NN). To be able to use a model in a safety-critical application, such as medical image recognition or autonomous driving, it is important to be able to quantify the uncertainty in the predictions of the model, [Paleyes et al., 2020]. For NNs, there are numerous methods to quantify the uncertainty in the predictions, [Gawlikowski et al., 2021]. Here, the delta method, [Liero and Zwanzig, 2011, Malmström, 2021] is an example of such a method. It relies on a two-step procedure. Firstly, to compute the uncertainty of the parameters in the black-box model, and secondly to propagate, through linearization, the uncertainty in the parameters to uncertainty in the output. Hence, it is a method based on identifying a distribution for the parameters of the black-box model. The method shares similarities to the Laplacian approximation of Bayesian NNs, [Immer et al., 2021]. Broadly speaking, methods to quantify the uncertainty in the predictions can be separated into two categories. The first category of methods is based on creating ensembles of predictions, [Gal and Ghahramani, 2016, Lakshminarayanan et al., 2017, Malmström, 2021]. The second category of methods is based on learning the uncertainty by modifying the NN structure or the cost function, [Blundell et al., 2015, Izmailov et al., 2021, Kendall and Gal, 2017].

For black-box models, such as NNs, a common design choice is to use an overparameterized model to guarantee that the model is flexible enough to describe the true system. In the literature, the problem of quantifying the uncertainty for overparameterized models has previously been studied, see [Hjalmarsson and Gustafsson, 1995, Pintelon et al., 1996, Stoica and Marzetta, 2001, Stoica and Söderström, 1989]. For example, independently of how a system has been overparameterized, [Pintelon et al., 1996] shows that the uncertainty in the learned (estimated) parameters is the same. In this paper, the contribution is to show, that using the delta method to compute the uncertainty in the prediction of a black-box model, the uncertainty is always larger for an overparameterized model compared to the calculated uncertainty from a model with minimum flexibility required to describe the true system. However, the calculated uncertainty in the prediction is the same if the added parameters of the overparameterized model do not add flexibility to the model.

⋆ This work is supported by Sweden's innovation agency, Vinnova, through project iQDeep (project number 2018-02700).

2. PROBLEM FORMULATION

This paper will consider regression problem with least-squares loss function, and investigate how the use of overparameterization affects the prediction uncertainty calculated by the delta method.

2.1 Signal model and likelihood

Consider a mapping between an input $\mathbf{x}_n \in \mathbb{R}^{n_x}$ to some output $f^*(\mathbf{x}_n) \in \mathbb{R}$. Here n_x is the dimension of the input. A parametric black-box model $f(\mathbf{x}_n; \boldsymbol{\theta})$ is used to model this relationship, where $\boldsymbol{\theta} \in \mathbb{R}^{n_{\boldsymbol{\theta}}}$ is the model parameters. Assume a scalar measurement $y_n \in \mathbb{R}$ of the output $f^*(\mathbf{x}_n)$ given by

$$y_n = f^*(\mathbf{x}_n) + e_n, \tag{1}$$

where e_n is i.i.d. measurement noise. The parameters of the black-box model are learned from the measurements y_n by minimizing a loss function $V_N(\boldsymbol{\theta})$, i.e.,

10.1016/j.ifacol.2023.10.077

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} V_N(\boldsymbol{\theta}). \tag{2}$$

The least-squares loss function

$$V_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} ||y_n - f(\mathbf{x}_n;\boldsymbol{\theta})||^2 \tag{3}$$

is a common choice for regression problems. Here N denotes the number of data points. If the noise e_n has a Gaussian distribution, and loss function (3) is used, the estimate in (2) is the maximum likelihood estimate (MLE).

2.2 Neural network model structure

A fully connected NN with L layers can be written as

$$\mathbf{h}^{(0)} = \mathbf{x}, \tag{4a}$$

$$\mathbf{a}^{(l+1)} = \begin{bmatrix}\mathbf{h}^{(l)} & 1\end{bmatrix}^\top W^{(l)}, \quad l = 0, \dots, L-1, \tag{4b}$$

$$\mathbf{h}^{(l)} = \sigma\big(\mathbf{a}^{(l)}\big), \quad l = 1, \dots, L-1, \tag{4c}$$

where $\sigma()$ denotes the activation function. The latent variable $\mathbf{a}^{(l)}$ containing the value of all the nodes in the l'th layer of the NN, and $\mathbf{h}^{(l)}$ denotes the transformation using the activation function of the values in all the nodes in the l'th layer of the NN. Collecting all the weights and biases included in the matrices $W^{(L)}, \dots, W^{(0)}$ into the parameter vector using the $\mathrm{Vec}(\cdot)$ function, i.e.,

$$\boldsymbol{\theta} \triangleq \begin{bmatrix}\mathrm{Vec}(W^{(L)})^\top & \dots & \mathrm{Vec}(W^{(0)})^\top\end{bmatrix}^\top, \tag{4d}$$

the NN can be written as a parametric model

$$f(\mathbf{x};\boldsymbol{\theta}) = \mathbf{a}^{(L)}. \tag{4e}$$

2.3 True system

Define a model set $\mathcal{M}^*$ as a collection of candidate models, [Ljung, 1999]. If the model set generated by the black-box model $f(\mathbf{x}_n;\boldsymbol{\theta})$ includes the true system $f^*(\mathbf{x}_n)$, then there exists a $\boldsymbol{\theta}_0$ such that

$$f(\mathbf{x}_n;\boldsymbol{\theta}_0) = f^*(\mathbf{x}_n). \tag{5}$$

In practice, it is difficult to guarantee that $\mathcal{M}^*$ includes the true system. However, if a very flexible model, such as a deep NN is used, it can be assumed that such a $\boldsymbol{\theta}_0$ exists, [Liang and Srikant, 2017]. Hence, for these applications, it is sensible to assume that the model set includes the true system. For NN, there are some symmetries in the parametrization that makes $\boldsymbol{\theta}_0$ non-unique. These symmetries can be handled similarly to overparameterization, but a description of them is outside of the scope of this paper. Interested readers are referred to [Hwang and Ding, 1997].

If the true system is contained in the model set $\mathcal{M}^*$, an estimate of $\boldsymbol{\theta}$ found by solving (2) is, asymptotically in the number of data points, the MLE. This is true whether the noise is Gaussian distributed or not, [Ljung, 1999]. From here on, throughout this paper, it will be assumed that all model sets include the true system, regardless of if they are generated by an overparameterized model or not.

2.4 Parameter covariance

Under the assumption that the signal-to-noise ratio (SNR) tends to infinity, the MLE gives that the estimated parameters convergence in distribution to

$$\hat{\boldsymbol{\theta}}_N \sim \mathcal{N}(\boldsymbol{\theta}_0, P_N^{\boldsymbol{\theta}}). \tag{6}$$

Here $P_N^{\boldsymbol{\theta}}$ is the Cramér-Rao lower bound (CRLB). For the regression problem in (3), the CRLB is given by

$$P_N^{\boldsymbol{\theta}} = \lambda_N \big(\mathcal{I}^{\boldsymbol{\theta}}\big)^{-1}, \tag{7a}$$

$$\mathcal{I}^{\boldsymbol{\theta}} = \sum_{n=1}^{N} \mathcal{I}_n^{\boldsymbol{\theta}}, \tag{7b}$$

$$\mathcal{I}_n^{\boldsymbol{\theta}} = E\big[\psi(\mathbf{x}_n;\boldsymbol{\theta})\psi^\top(\mathbf{x}_n;\boldsymbol{\theta})\big], \tag{7c}$$

where $\mathcal{I}^{\boldsymbol{\theta}}$ denotes the information matrix, the prediction error and the prediction error variance are given by

$$\epsilon(\mathbf{x}_n;\boldsymbol{\theta}) = y_n - f(\mathbf{x}_n;\boldsymbol{\theta}), \tag{8a}$$

$$\lambda_N = \frac{1}{N}\sum_{n=1}^{N} \epsilon^2(\mathbf{x}_n;\boldsymbol{\theta}), \tag{8b}$$

and the derivative of the prediction error with respect to the parameters $\boldsymbol{\theta}$

$$\psi(\mathbf{x}_n;\boldsymbol{\theta}) = -\frac{\partial}{\partial\boldsymbol{\theta}}\epsilon(\mathbf{x}_n;\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} f(\mathbf{x}_n;\boldsymbol{\theta}). \tag{9}$$

2.5 Error propagation using the delta method

The delta method [Liero and Zwanzig, 2011], is a method to propagate the uncertainty in the parameter to uncertainty in the prediction using a linearization of the model. That is, the covariance of the prediction is given by

$$P_N^f(\mathbf{x}) \triangleq \mathrm{var}\big(f(\mathbf{x};\hat{\boldsymbol{\theta}}_N)\big) = \psi^\top(\mathbf{x};\boldsymbol{\theta}) P_N^{\boldsymbol{\theta}} \psi(\mathbf{x};\boldsymbol{\theta}), \tag{10}$$

for any input $\mathbf{x}$. It is based on the observation that, if the parameters are Gaussian distributed, close to the parameters $\hat{\boldsymbol{\theta}}_N$, the model can accurately be represented by a linearized local approximation around $f(\mathbf{x};\hat{\boldsymbol{\theta}}_N)$, Liero and Zwanzig [2011]. In system identification, this is a standard method to project the uncertainty of the parameters onto the prediction, see [Chryssoloiuris et al., 1996, Hwang and Ding, 1997, Papadopoulos et al., 2001, Rivals and Personnaz, 2000]. Then, the uncertainty in the prediction is given by the norm of the projection, [Hjalmarsson and Martensson, 2010].

This paper will study how the prediction uncertainty given by (10) is affected by using an overparameterized model. To do so, overparameterized models will be separated into two categories:

(i) Models where the redundant parameters do not add any flexibility, e.g., when some elements in $\mathbf{a}^{(l)}$ can be written as a linear combination of the other elements of $\mathbf{a}^{(l)}$. These will be referred to as redundant parameters of Category 1.
(ii) Models where the redundant parameters add flexibility, e.g., when more hidden nodes than necessary are used for an NN. This will be referred to as redundant parameters of Category 2.

3. OVERPARAMETIZED MODELS

The delta method is based on a linear approximation of a nonlinear model. Hence, to analyze how (10) is affected by overparameterization, consider the linearization

$$f(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\varphi}^\top(\mathbf{x})\boldsymbol{\theta} + d, \tag{11}$$

where $\boldsymbol{\varphi}(\mathbf{x}) = \psi(\mathbf{x};\boldsymbol{\theta})$, and d is the difference between the value of the function and the linear approximation

at the linearization point. However, d will not influence the uncertainty in the prediction calculated by the delta method, hence assume $d = 0$. Note that for a model linear in the parameters, the information matrix can be written as

$$\mathcal{I}^{\boldsymbol{\theta}} = \Phi_N \Phi_N^\top, \tag{12a}$$

where

$$\Phi_N = [\boldsymbol{\varphi}(\mathbf{x}_1) \ \cdots \ \boldsymbol{\varphi}(\mathbf{x}_N)]. \tag{12b}$$

A model of minimum flexibility but still flexible enough to describe the true system can be referred to as a canonical model. To make a distinction between whether the model is canonical or overparameterized, a subindex c is added here.

3.1 Models with redundant parameters of Category 1

As an example that the true input-output relationship is given by

$$f^*(x) = 1 + 6x + x^2. \tag{13}$$

Then one canonical model is given by the regressor $\boldsymbol{\varphi}_c^\top(x) = [1, x, x^2]$, while $\boldsymbol{\varphi}(x)^\top = [1, x, x+1, x^2]$ is an overparameterized model with redundant parameters of Category 1. Hence, there exists some transformation $T(\boldsymbol{\theta}) = \boldsymbol{\theta}_c$ which transforms the parameters of the overparameterized to parameters of the canonical representation model. In the linear case,

$$\boldsymbol{\theta}_c = T\boldsymbol{\theta}. \tag{14}$$

Here $T \in \mathbb{R}^{n_{\boldsymbol{\theta}_c} \times n_{\boldsymbol{\theta}}}$ is a transformation matrix where $n_{\boldsymbol{\theta}_c}$ is the number of parameters in the canonical model which is smaller than or equal to $n_{\boldsymbol{\theta}}$. The case where $n_{\boldsymbol{\theta}_c} = n_{\boldsymbol{\theta}}$ represents when there exist some symmetries in the model, e.g., changing the ordering of the nodes in (4).

For (7) to hold, it is assumed that the information matrix $\mathcal{I}^{\boldsymbol{\theta}}$ is invertible. That is, the data is informative enough, and the model is not overparameterized. Assume that the data is informative enough with respect to the model set generated by the canonical model. For an overparameterized model with redundant parameters of Category 1, where the information matrix is singular, the inverse is replaced by a Moore-Penrose pseudo-inverse denoted with the superscript $+$, [Stoica and Marzetta, 2001], i.e.,

$$P_N^{\boldsymbol{\theta}} = \lambda_N \big(\mathcal{I}^{\boldsymbol{\theta}}\big)^+, \tag{15}$$

which also works when there exists a null space in the parameter space.

Theorem 1. Consider a canonical model of the true system $f^*(\mathbf{x}) = \boldsymbol{\varphi}_c^\top(\mathbf{x})\boldsymbol{\theta}_{c,0}$. Form a new model with more parameters $\boldsymbol{\varphi}^\top(\mathbf{x})\boldsymbol{\theta}$ using the (wide) transformation matrix T with full row rank. Then, the models estimated by (2) will give identical uncertainty in the prediction independently of the choice of T, i.e.,

$$\boldsymbol{\varphi}_c^\top(\mathbf{x}_m) P_N^{\boldsymbol{\theta}_c} \boldsymbol{\varphi}_c(\mathbf{x}_m) = \boldsymbol{\varphi}^\top(\mathbf{x}_m) P_N^{\boldsymbol{\theta}} \boldsymbol{\varphi}(\mathbf{x}_m). \tag{16}$$

Proof. With the transformation given in (14), the regressors for the overparameterized model can be written as $\boldsymbol{\varphi}^\top(\mathbf{x}_m) = \boldsymbol{\varphi}_c^\top(\mathbf{x}_m)T$. Hence, the information matrix for the overparameterized model can be written as

$$\mathcal{I}^{\boldsymbol{\theta}} = T^\top \mathcal{I}^{\boldsymbol{\theta}_c} T. \tag{17}$$

Recall that, if A has full column rank, B has full row rank and C is invertible the pseudo-inverse of their product is

$$(ACB)^+ = B^+ C^{-1} A^+. \tag{18}$$

Let I_r denote the identity matrix of size r. Using that $TT^+ = I_{n_{\boldsymbol{\theta}_c}}$, the uncertainty in the prediction for the canonical model can be written as

$$\begin{aligned}
&\boldsymbol{\varphi}_c^\top(\mathbf{x}_m) P_N^{\boldsymbol{\theta}_c} \boldsymbol{\varphi}_c(\mathbf{x}_m) = && \text{(19a)}\\
&= \lambda_N \boldsymbol{\varphi}_c^\top(\mathbf{x}_m) T T^+ \big(\mathcal{I}_c^{\boldsymbol{\theta}}\big)^{-1} \big(T^\top\big)^+ T^\top \boldsymbol{\varphi}_c(\mathbf{x}_m) && \text{(19b)}\\
&= \lambda_N \boldsymbol{\varphi}^\top(\mathbf{x}_m) \big(T^\top \mathcal{I}^{\boldsymbol{\theta}_c} T\big)^+ \boldsymbol{\varphi}(\mathbf{x}_m) && \text{(19c)}\\
&= \boldsymbol{\varphi}^\top(\mathbf{x}_m) P_N^{\boldsymbol{\theta}} \boldsymbol{\varphi}(\mathbf{x}_m) && \text{(19d)}
\end{aligned}$$

This is identical to the uncertainty from (10) which was calculated using an overparameterized model with redundant parameters of Category 1.

□

Similar results have been presented in [Stoica and Marzetta, 2001]. There it is shown that even though the information matrix for the parameters is singular, a projection of the matrix can give an invertible information matrix. In [Pintelon et al., 1996] it is shown that the CRLB is independent of the given overparameterization. The main difference in Theorem 1, is that it is the uncertainty in the prediction that is considered where equivalence is shown to a canonical model. Thereby, it is shown that the effect of the null space in $\mathcal{I}^{\boldsymbol{\theta}}$ can be neglected. This is a result of the structure from having the same transformation when computing the parameter covariance and propagating the uncertainty.

3.2 Models with redundant parameters of Category 2

Once again, consider the problem to identify a model for the system in (13), but now, the overparameterized model has parameters that add unnecessary flexibility, i.e., redundant parameters of Category 2. For example if the regressor $\boldsymbol{\varphi}(x) = [1, x, x^2, x^3]^\top$ is used. If the model has redundant parameters of Category 2, the information matrix for the overparameterized model is likely to be invertible.

Theorem 2. Consider the linear model (11) and assume that the true system can be described by a canonical model with less flexibility. If both models are estimated using (2) where $n_{\boldsymbol{\theta}}$ is fixed and $n_{\boldsymbol{\theta}} \ll N$, then as the SNR goes to infinity, the uncertainty in the prediction (10) for a canonical model is smaller compared to that of an overparameterized model, i.e.,

$$\boldsymbol{\varphi}_c^\top(\mathbf{x}_m) P_N^{\boldsymbol{\theta}_c} \boldsymbol{\varphi}_c(\mathbf{x}_m) < \boldsymbol{\varphi}^\top(\mathbf{x}_m) P_N^{\boldsymbol{\theta}} \boldsymbol{\varphi}(\mathbf{x}_m). \tag{20}$$

Proof. Without loss of generality, the overparameterized model can be written in terms of the canonical model such as

$$\begin{aligned}
\boldsymbol{\varphi}^\top(\mathbf{x}_m) &= \begin{bmatrix}\boldsymbol{\varphi}_c^\top(\mathbf{x}_m) & \boldsymbol{\varphi}_o^\top(\mathbf{x}_m)\end{bmatrix}, && \text{(21a)}\\
\Phi_N^\top &= \begin{bmatrix}\Phi_{c,N}^\top & \Phi_{o,N}^\top\end{bmatrix}, && \text{(21b)}\\
\Phi_N \Phi_N^\top &= \begin{bmatrix}\Phi_{c,N}\Phi_{c,N}^\top & \Phi_{c,N}\Phi_{o,N}^\top \\ \Phi_{o,N}\Phi_{c,N}^\top & \Phi_{o,N}\Phi_{o,N}^\top\end{bmatrix}, && \text{(21c)}
\end{aligned}$$

where $\boldsymbol{\varphi}_o^\top(\mathbf{x}_m)$ and $\Phi_{o,N}^\top$ correspond to the parameters added in the overparameterized model. Using the blockwise inverse the parameter inverse of the information

matrix for the parameters of the overparameterized model is

$$\left(\Phi_N\Phi_N^\top\right)^{-1}=\begin{bmatrix}\left(\Phi_{c,N}\Phi_{c,N}^\top\right)^{-1}+K^\top R_oK & -K^\top R_o\\ -R_oK & R_o\end{bmatrix}, \quad (22a)$$

where

$$R_o=\left(\Phi_{o,N}(I_N-R_c)\Phi_{o,N}^\top\right)^{-1}, \quad (22b)$$

$$R_c=\Phi_{c,N}^\top\left(\Phi_{c,N}\Phi_{c,N}^\top\right)^{-1}\Phi_{c,N}, \quad (22c)$$

$$K=\Phi_{o,N}\Phi_{c,N}^\top\left(\Phi_{c,N}\Phi_{c,N}^\top\right)^{-1}. \quad (22d)$$

Since $n_{\boldsymbol{\theta}} \ll N$ and both the model set of the canonical and overparameterized model include the true system, the prediction error variance is asymptotically the same and equal to the variance of the noise λ_0, [Ljung, 1999]. Then the variance of the overparameterized model can be written as

$$\begin{aligned}\lambda_0\boldsymbol{\varphi}^\top(\mathbf{x}_m)\left(\Phi_N\Phi_N^\top\right)^{-1}\boldsymbol{\varphi}(\mathbf{x}_m)&=\lambda_0\boldsymbol{\varphi}_c^\top(\mathbf{x}_m)\left(\Phi_{c,N}\Phi_{c,N}^\top\right)^{-1}\boldsymbol{\varphi}_c(\mathbf{x}_m)\\ &+\lambda_0\boldsymbol{\varphi}_c^\top(\mathbf{x}_m)K^\top R_oK\boldsymbol{\varphi}_c(\mathbf{x}_m)-\lambda_0\boldsymbol{\varphi}_c^\top(\mathbf{x}_m)K^\top R_o\boldsymbol{\varphi}_o(\mathbf{x}_m)\\ &-\lambda_0\boldsymbol{\varphi}_o^\top(\mathbf{x}_m)R_oK\boldsymbol{\varphi}_c(\mathbf{x}_m)+\lambda_0\boldsymbol{\varphi}_o^\top(\mathbf{x}_m)R_o\boldsymbol{\varphi}_o(\mathbf{x}_m). \quad (23)\end{aligned}$$

Here (23) can be seen as a quadratic optimization problem where the regressor that adds flexibility to the model $\boldsymbol{\varphi}_o^\top(\mathbf{x}_m)$ and $\Phi_{o,N}$ are free to choose. Write

$$\boldsymbol{\chi}^\top=[\boldsymbol{\varphi}_c^\top(\mathbf{x}_m)K^\top,\boldsymbol{\varphi}_o^\top(\mathbf{x}_m)], \quad (24a)$$

then the optimization problem becomes

$$\min_{\boldsymbol{\chi}}\ \lambda_0\boldsymbol{\varphi}_c^\top(\mathbf{x}_m)\left(\Phi_{c,N}\Phi_{c,N}^\top\right)^{-1}\boldsymbol{\varphi}_c(\mathbf{x}_m)+\lambda_0\boldsymbol{\chi}^\top Q\boldsymbol{\chi}, \quad (24b)$$

where

$$Q=\begin{bmatrix}R_o & -R_o\\ -R_o & R_o\end{bmatrix}. \quad (24c)$$

Since R_o is invertible by definition, Q is a positive semi-definite matrix, the minimum is found when $\boldsymbol{\chi}^\top Q\boldsymbol{\chi}=0$. In order to obtain a non-trivial solution, one has to choose $\boldsymbol{\varphi}_c^\top(\mathbf{x}_m)K^\top=\boldsymbol{\varphi}_o^\top(\mathbf{x}_m)$, i.e., the added flexibility has to be written as a linear combination of the canonical model. Hence, adding flexibility to the overparameterized model will always increase (24b). This concludes that the calculated uncertainty in the prediction for an overparameterized model with redundant parameters of Category 2 will be larger compared to that of the canonical model.

□

Note that choosing $\boldsymbol{\varphi}_c^\top(\mathbf{x}_m)K^\top=\boldsymbol{\varphi}_o^\top(\mathbf{x}_m)$ would result in that neither (21c) nor R_o is invertible, i.e., the same setting as Theorem 1, where $T=K^\top$. Hence equality in (20) can only be obtained when the redundant parameters are of Category 1.

By formulating a model selection problem, for nested model structures, similar results as Theorem 2 have been shown in [Hjalmarsson and Gustafsson, 1995] where the calculated uncertainty is higher for the larger model. That the uncertainty in the prediction is higher for an overparameterized model compared to a canonical model, is also the premise for using model selection algorithms, such as the Bayesian information criteria (BIC), [Schwarz, 1978]. This paper provides insights into why the calculated uncertainty has to be strictly larger for the overparameterized model, and how the result generalizes to the case with redundant parameters of Category 1. That is since the minimum of the minimization problem in (24b) is obtained when the added regressor can be written as a linear combination of the regressor of the canonical model, added flexibility must increase the calculated uncertainty in the prediction.

The results of Theorem 1 and Theorem 2 can be summarised as for an overparameterized model, the calculated uncertainty from using the delta method will always be larger compared to a canonical model. However, if the additional parameters do not add any flexibility, the uncertainty in the prediction is the same. In practice, by observing structure of the chosen model, it is hard to distinguish between the two aforementioned categories of overparameterization. However, after quantifying the uncertainty, if the information matrix $\mathcal{I}^{\boldsymbol{\theta}}$ is rank deficient it would indicate that it is likely that the model might have redundant parameters of Category 1.

3.3 Nonlinear models

The main idea of the delta method is a two-step linearization. Firstly, to compute the parameter uncertainty, and secondly to propagate the uncertainty to the output of the model. Hence, the delta method gives a linear approximation of a nonlinear model. Asymptotically in the number of data points, close to the MLE, a good approximation for many nonlinear models is given by a linear approximation, [Enqvist, 2005, Ljung and Glad, 1994, Nocedal and Wright, 2006]. Consequently, using (10) to calculate the uncertainty in the prediction for these nonlinear models, the result of Theorem 1 and Theorem 2 should hold asymptotically close to the estimated parameters since the delta method uses a linear approximation of the nonlinear model.

4. NUMERICAL EXAMPLES

A simulation study will be used to validate the results from Theorem 1 and Theorem 2. The true system under consideration is how the normalized traction force depends on the wheel slip referred to as the *magic formula tire model*, [Pacejka and Besselink, 1997],

$$f^*(x)=D\sin\left(C\arctan\left(Bx-E(Bx-\arctan(Bx))\right)\right). \quad (25)$$

It is exactly modeled by an NN with two layers and two nodes in the hidden layer [Malmström, 2021]. Hence the true system is included in the model set. For the simulation $B=14$, $C=0.1$, $D=0.6$, $E=-0.2$ in (25), 200 mesurements are generated where $x\sim\mathcal{U}[-0.6,0.6]$, and $e\sim\mathcal{N}(0,0.01)$.

4.1 Models with redundant parameters of Category 1

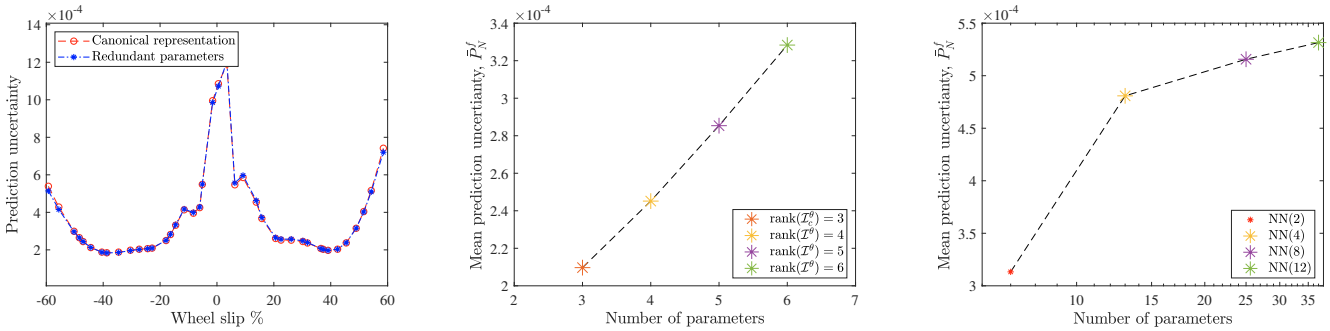
A nonlinear canonical model given by

$$f_c(x,\boldsymbol{\theta}^c)=\boldsymbol{\theta}_1^c\sigma(\boldsymbol{\theta}_2^cx+\boldsymbol{\theta}_3^c)+\boldsymbol{\theta}_4^c\sigma(\boldsymbol{\theta}_5^cx+\boldsymbol{\theta}_6^c)+\boldsymbol{\theta}_7^c, \quad (26)$$

and an example of an overparameterized model with redundant parameters of Category 1 is e.g.,

$$\begin{aligned}f(x,\boldsymbol{\theta})=&\boldsymbol{\theta}_1\sigma(\boldsymbol{\theta}_2(x+1)+\boldsymbol{\theta}_3(2-x)+5\boldsymbol{\theta}_4)\\ &+\boldsymbol{\theta}_5\sigma(\boldsymbol{\theta}_6x+\boldsymbol{\theta}_7)+\boldsymbol{\theta}_8. \quad (27)\end{aligned}$$

The parameters of these models are estimated using (2), and then the uncertainty in the parameters and prediction is computed using (10). From Fig. 1a one can conclude that the uncertainty of the overparameterized model with redundant parameters of Category 1 is the same as the uncertainty of the canonical one.



(a) Nonlinear model with redundant parameters Category 1.

(b) Linear model with redundant parameters Category 2.

(c) Nonlinear model with redundant parameters Category 2.

Fig. 1. Calculated uncertainty in the prediction for a canonical and an overparameterized model using the delta method. In (a) the model is nonlinear and the overparameterized model has redundant parameters of Category 1. While in (b) the model is linear in its parameters and the overparameterization has redundant parameters of Category 2 which add flexibility to the model. In (c) the model is a two-layer NN with an increasing number of nodes in the hidden layer, i.e., it is nonlinear and the overparameterization with redundant parameters of Category 2. The simulation data is generated by (25).

4.2 Models with redundant parameters of Category 2

A canonical model linear in its parameters is given by

$$\varphi_c(x)=[\sigma(-40x+0.0061), \sigma(-6.8x+0.0036), 1]^\top. \quad (28)$$

And a model with more flexibility but with redundant parameters of Category 2 is defined by

$$\varphi^\top(x) = [\varphi_c^\top(x), \varphi_{o_j}^\top(x)], \quad (29a)$$

$$\varphi_{o_j}^\top(x) = [\sigma(W_{o_j}x + b_{o_j}), \varphi_{o_{j-1},m}^\top], \quad (29b)$$

where $o_j = 1, ..., 3$, and $W_{o_j} = b_{o_j} = o_j$.

A nonlinear canonical model is given by a fully connected two-layer NN with two nodes in the hidden layer and sigmoid as an activation function, i.e., (4) with $L = 2$. Increasing the number of nodes in the hidden overparameterized model with redundant parameters of Category 2 can be obtained.

Once again the parameters are estimated using (2) and the uncertainty in the prediction is calculated using (10), both for the nonlinear model and the model linear in its parameters. For each model, the mean uncertainty is computed for N_v validation data points according to

$$\bar{P}_N^f = \frac{1}{N_v}\sum_{i=1}^{N_v} P_N^f(\mathbf{x}_i). \quad (30)$$

From Fig. 1b and Fig. 1c, Theorem 2 can be verified as the models with redundant parameters of Category 2 have a higher mean prediction uncertainty compared to the canonical model.

5. CONCLUSION

When calculating the uncertainty in the prediction using the delta method, for an overparameterized model, the prediction uncertainty will always be larger or equal compared to one of the models with minimum flexibility that still can describe the true system, i.e., a canonical model. Hence, the conclusion is that even though the model is overparameterized, the uncertainty in the prediction calculated by the delta method is not underestimated, i.e., the uncertainty in the prediction is not too low. The delta method relies on linearizations, hence asymptotically in the number of data points, the results from Theorem 1 and Theorem 2 apply locally around the estimated parameters for some nonlinear models. Section 4 provides an example of two such nonlinear models.

A future research direction could be to investigate if the result also holds for larger black-box models such as NNs used for image classification.

REFERENCES

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight Uncertainty in Neural Networks. In *Proc. of the 32nd Int. Conf. on Mach. Learn. (ICML)*., 1613–1622. Lille, France. 6–11 Jul.

Chryssoloiuris, G., Lee, M., and Ramsey, A. (1996). Confidence interval prediction for neural network models. In *IEEE Trans. Neural Netw.*, volume 7, 229–232.

Enqvist, M. (2005). *Linear models of nonlinear systems*. Ph.D. thesis, Dept. Elect. Eng., Linköping University, Linköping, Sweden.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proc. of the 33td Int. Conf. on Mach. Learn. (ICML)*., 1050–1059. New York, NY, USA. 20–22 Jun.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Muhammad, S., Wen Yang, Richard, B., and Xiao Xiang, Z. (2021). A survey of uncertainty in deep neural networks. In *arXiv preprint arXiv:2107.03342*.

Hjalmarsson, H. and Gustafsson, F. (1995). Composite modeling of transfer functions. In *IEEE Trans. Autom. Control*, volume 40, 820–832. IEEE.

Hjalmarsson, H. and Martensson, J. (2010). A geometric approach to variance analysis in system identification. In *IEEE Trans. Autom. Control*, volume 56, 983–997. IEEE.

Hwang, J.T.G. and Ding, A.A. (1997). Prediction Intervals for Artificial Neural Networks. In *J. Am. Stat. Assoc. (JSTOR)*, volume 92, 748–757. Taylor & Francis.

Immer, A., Korzepa, M., and Bauer, M. (2021). Improving predictions of bayesian neural nets via local linearization. In *Proc. of 24nd Int. Conf. on Artificial Intell.*

and Statistics. (AISTATS), 703–711. PMLR, San Diego, CA, USA. 13-15 Apr.

Izmailov, P., Nicholson, P., Lotfi, S., and Wilson, A.G. (2021). Dangers of bayesian model averaging under covariate shift. In *Adv. in Neural Inf. Process. Syst. (NIPS) 35*, volume 34. Virtual.

Kendall, A. and Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Adv. in Neural Inf. Process. Syst. (NIPS) 31*, 5574–5584. Curran Associates, Inc. Long Beach, CA, USA, 4–9 Dec.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Adv. in Neural Inf. Process. Syst. (NIPS) 31*. Curran Associates, Inc. Long Beach, CA, USA, 4–9 Dec.

Liang, S. and Srikant, R. (2017). Why deep neural networks for function approximation? In *Proc. of IEEE Int. Conf. on Robot. and Autom. (ICRA)*. IEEE, Singapore, Singapore. 19 May–3 June.

Liero, H. and Zwanzig, S. (2011). *Introduction to the theory of statistical inference*. Chapman and Hall CRC Texts in Statistical Science, Boca Raton, FL, USA.

Ljung, L. (1999). *System identification: theory for the user (2nd edition)*. PTR Prentice Hall: Upper Saddle River, NJ, USA.

Ljung, L. and Glad, T. (1994). *Modeling of dynamic systems*. Prentice-Hall, Englewood Cliffs,NJ, USA.

Malmström, M. (2021). *Uncertainties in Neural Networks A System Identification Approach*. Licentiate thesis, Dept. Elect. Eng., Linköping University, Linköping, Sweden.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media,: New York, NY, USA.

Pacejka, H. and Besselink, I. (1997). Magic formula tyre model with transient properties. *Veh. syst. dynamics-Int. J. of Veh. Mechanics and Mobility*, 27(S1), 234–249.

Paleyes, A., Urma, R.G., and Lawrence, N.D. (2020). Challenges in deploying machine learning: a survey of case studies. In *Adv. in Neural Inf. Process. Syst. (NIPS) 34 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)*, volume 33. Virtual.

Papadopoulos, G., Edwards, P., and Murray, A. (2001). Confidence estimation methods for neural networks: a practical comparison. In *IEEE Trans. Neural Netw.*, volume 12, 1278–1287.

Pintelon, R., Schoukens, J., McKelvey, T., and Rolain, Y. (1996). Minimum variance bounds for overparameterized models. In *IEEE Trans. Autom. Control*, volume 41, 719–720.

Rivals, I. and Personnaz, L. (2000). Construction of confidence intervals for neural networks based on least squares estimation. In *Elsevier J. Neural Netw.*, volume 13, 463–484. Elsevier.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. of Stat.*, 6(2), 461–464.

Stoica, P. and Marzetta, T.L. (2001). Parameter estimation problems with singular information matrices. In *IEEE Trans. Signal Process.*, volume 49, 87–90.

Stoica, P. and Söderström, T. (1989). On reparametrization of loss functions used in estimation and the invariance principle. In *Signal processing*, volume 17, 383–387. Elsevier.