QUANTIFICATION VS. REDUCTION: ON EVALUATING REGRESSION UNCERTAINTY

Domokos M. Kelen[*†] **Ádám Jung**[†] **András A. Benczúr**[†‡]

ABSTRACT

Uncertainty quantification (UQ) methods for regression are frequently judged based on improvements measured in negative log-likelihood (NLL). In this work, we question the practice of relying too heavily on NLL, arguing that typical evaluations can conflate *better quantifying* predictive uncertainty with simply *reducing* it. We do so by studying how the uncertainty of various distributional parameters affects NLL scoring. In particular, we demonstrate how the error of the mean materializes as uncertainty, and how the uncertainty of the variance has almost no effect on scores. Our results question how much of the reported progress is due to decreasing, rather than accurately representing, uncertainty, highlighting the need for additional metrics and protocols that disentangle these two factors.

1 INTRODUCTION

Uncertainty quantification (UQ) methods provide a framework for estimating the confidence of predictive models. Over the years, numerous UQ techniques have been proposed and evaluated based on improvements in negative log-likelihood (NLL). While NLL is widely regarded as a robust metric such as in the UCI Regression Uncertainty Benchmark (Hernández-Lobato & Adams, 2015) due to its strictly proper scoring properties, we demonstrate that its ability of assessing uncertainty quantification is confounded by improvements in overall predictive performance in regression.

A key concern in relying too heavily on NLL for evaluating UQ methods is that it does not strictly measure the accuracy of uncertainty quantification but rather reflects how well the predicted distribution aligns with observed outcomes. Through proving two propositions, we formally show how NLL can be reduced not just by genuinely refining uncertainty estimates, but also by merely improving overall prediction accuracy. Since NLL is inherently linked to the entropy of the predicted distribution, a narrower target distribution, resulting from reduced predictive errors, naturally leads to lower NLL scores. For our propositions to hold, ground truth can exist for the aleatoric uncertainty, unlike assumed in the related argument of (Bengs et al., 2023).

The problem is excerberated by the fact that many UQ methods incorporate elements resembling regularization techniques, which are specifically designed to decrease predictive uncertainty and improve generalization, yet might not lead to a more accurate quantification of uncertainty. This raises fundamental questions about the validity of using NLL as the primary evaluation metric for UQ methods, and underscores the necessity of additional metrics and evaluation protocols that disentangle the effects of improved predictive performance from better uncertainty quantification.

Our paper critically examines the limitations of NLL as a standalone metric for UQ evaluation. Specifically, we explore how epistemic and aleatoric uncertainty influence predictive distributions, and investigate the effects of accuracy in predicting parameters of the aleatoric uncertainty distribution. We demonstrate how better point predictions for the aleatoric mean μ lead to decreased predictive uncertainty during evaluation, and quantify the effects of uncertainty regarding the distribution of the aleatoric parameter σ^2. In particular, we show that NLL scores appear almost independent of the uncertainty in predicting σ^2, to the point of enabling the use of the global variance as an unconditional estimate.

[*]Emails: {kdomokos, jungadam, benczur}@info.ilab.sztaki.hu
[†]HUN-REN SZTAKI [‡]Széchenyi University, Győr, Hungary

By doing so, we demonstrate the potentially highly misleading nature of traditional NLL evaluation. Overall, our aim is to encourage the use of rigorous assessment of UQ techniques, and to encourage further research into selecting appropriate metrics and evaluation protocols.

2 BACKGROUND

Typically, the data we record originates from a real-world process that is only partially observed. In regression, the target value (i.e., the value to be predicted) can be modeled as a random variable $Y : \Omega \to \mathbb{R}$, with each $\omega \in \Omega$ representing the true state of the world. Although ω itself is unobservable, we can measure feature variables $X : \Omega \to \mathbb{R}^m$, which offer partial information about ω. The classic regression objective is then to approximate the conditional expectation $\mathbb{E}[Y \mid X = x, D]$, where D is a sample of past (X, Y) values, while UQ is concerned with the distribution $p(Y \mid X = x, D)$.

2.1 ALEATORIC VS. EPISTEMIC UNCERTAINTY

Different sources of uncertainty can be distinguished (Hüllermeier & Waegeman, 2021). *Epistemic uncertainty* arises from our inability to pinpoint the optimal model based on limited data, and can be reduced by collecting more data (i.e., expanding D). This is typically formalized as examining the conditional distribution of model parameters given the data, for example in Bayesian Neural Networks (Blundell et al., 2015; Gawlikowski et al., 2023). *Aleatoric uncertainty*, by contrast, stems from the fact that knowing X may not fully determine Y, as some component of Y is independent of X. As a result, it cannot be reduced without redefining X or Y.

2.2 NEGATIVE LOG-LIKELIHOOD, PROPER SCORING RULES, AND DIFFERENTIAL ENTROPY

Negative log-likelihood (NLL) is commonly used to measure how closely a predicted probability distribution matches observed data. Formally, it is defined as

$$-\mathbb{E}\big[\log \tilde{q}_\theta(Y \mid X)\big],$$

where $\tilde{q}_\theta(Y \mid X)$ denotes the model's predicted distribution for the target variable. NLL is a *strictly proper scoring rule* (Gneiting & Raftery, 2007), meaning it is minimized exactly when the predicted distribution matches the true data-generating distribution, thereby encouraging accurate estimates.

When the predicted distribution is identical to the ground truth, the expected NLL reduces to the differential entropy $-\mathbb{E}[\log p(Y \mid X)]$. Thus, predicting a distribution with lower entropy naturally leads to lower range of NLL values.

3 RELATED WORK

In UQ, the majority of methods focus on quantifying epistemic uncertainty, for example the BNN (Blundell et al., 2015), or its many enhancements (Gawlikowski et al., 2023). Goodness of regression uncertainty approaches is usually measured on the UCI Benchmark (Hernández-Lobato & Adams, 2015), with most papers reporting RMSE and (negative) log-likelihood (NLL) results.

In Bengs et al. (2023), the possibility of second-order scoring rules is investigated, i.e., scoring rules for epistemic uncertainty. Their results indicate that such scoring rules most likely cannot exists, thereby also questioning the validity of using first-order scoring rules such as NLL. However, their result hinges on the fact that no ground truth exists for the aleatoric uncertainty distribution. Our results venture further, highlighting the misleading nature of NLL evaluation even in cases where we assume nonexistent or known aleatoric uncertanty.

NLL conflating improvements of accuracy and UQ is not a new idea, with multiple recent publications describing the problem. However, both Sluijterman et al. (2024) and Kristoffersson Lind et al. (2024) approach the matter by demonstrating the effect through brief numeric evaluations. We provide a more formal analysis, contrasting the problem with scoring rule theory, aleatoric and epistemic uncertainty, and studying the effects of the uncertainty of different aleatoric parameters.

4 UNCERTAINTY QUANTIFICATION VS. REDUCTION

In this section, we study the exact ways in which comparing predicted distributions against the data using NLL can conflate better quantifying uncertainty with simply reducing uncertainty.

4.1 SIMPLIFIED DETERMINISTIC SETTING

To illustrate our point, we begin with a highly simplified problem setup. Suppose there is no inherent (aleatoric) noise in Y given X, so that

$$Y = \mathbb{E}[Y \mid X]. \tag{1}$$

In other words, Y is fully determined by X. The model's core task is then simply to learn the function

$$\mu(x) = \mathbb{E}[Y \mid X = x]. \tag{2}$$

A typical regression model might produce a point estimate $\tilde{\mu}(x;\theta)$ that approximates $\mu(x)$. An uncertainty-aware model, however, outputs a distribution over possible values for $\mu(x)$. For example, a Bayesian neural network (BNN) might represent the predicted distribution of $\mu(x)$ by producing different samples with each evaluation, interpreted as samples from the distribution.

Let us denote the model's predicted distribution over the true parameter $\mu(x)$ as

$$\tilde{q}_\theta\big(\mu(x) \mid X = x\big). \tag{3}$$

A standard evaluation via negative log-likelihood (NLL) is then

$$-\mathbb{E}\Big[\log\, \tilde{q}_\theta\big(Y \mid X\big)\Big] \approx -\frac{1}{n}\sum_{i=1}^{n} \log\, \tilde{q}_\theta\big(y_i \mid x_i\big), \tag{4}$$

where (x_i, y_i) are samples from the ground-truth joint distribution of (X, Y).

The usual argument for NLL-based evaluation is that NLL is a strictly proper scoring rule, which means it is minimized precisely when the predicted distribution matches the true distribution. Under this argument, the predictor is incentivized to align the predicted distribution with the ground-truth.

However, this assumes that the *ground-truth distribution* being approximated is fixed, which in our setup is *not* the case. As we have assumed aleatoric uncertainty in Y to be zero, the only distribution left to approximate is that of the model's own errors, not a fixed noise distribution. If the model's errors decrease, the associated "true" distribution of those errors shrinks; if the model's errors increase, it broadens. In both scenarios, uncertainty can be deemed accurately quantified if the predicted uncertainty distribution matches the (changing) error distribution.

A narrower target distribution inherently corresponds to a lower range of NLL values: the ideal NLL is always equal to the differential entropy of the target distribution, and it is known that $H(\alpha X) = H(X) + \log|\alpha|$. Thus, a more accurate (lower-error) model with a less well-calibrated uncertainty estimate can outscore a less accurate model with a perfectly calibrated estimate. This conflates *reducing* predictive error with *quantifying* it more effectively.

4.1.1 GAUSSIAN EXAMPLE

In this section, we illustrate the problem through a concrete example. Again assume Equation (1), and simplifying even further, that the model posits a Gaussian density:

$$\tilde{q}_\theta\big(y \mid X = x\big) \;=\; \mathcal{N}\Big(y \,\Big|\, \tilde{\mu}(x;\theta),\; \tilde{\sigma}^2(x;\theta)\Big). \tag{5}$$

Denoting the residual as

$$\varepsilon(x) \;=\; Y - \tilde{\mu}(x;\theta), \tag{6}$$

the negative log-likelihood can be expressed as

$$-\,\mathbb{E}\big[\log \tilde{q}_\theta(Y \mid X)\big] \;=\; -\,\mathbb{E}\Big[\log\Big(\mathcal{N}\big(\varepsilon(X) \,\big|\, 0,\; \tilde{\sigma}^2(X;\theta)\big)\Big)\Big]. \tag{7}$$

Essentially, the predicted distribution is being compared against the residual distribution $\varepsilon(X)$. However, crucially, $\varepsilon(x)$ is itself defined by the model's chosen function $\tilde{\mu}(x;\theta)$. As $\tilde{\mu}(x;\theta)$ becomes more accurate, the distribution of $\varepsilon(x)$ tightens around zero; while if accuracy worsens, it broadens. In other words, the "target distribution" we compare against is not fixed, but depends on the quality of the model's point estimate.

Thus, the strictly proper scoring rule argument can be misleading in this setting: it implicitly assumes a fixed ground-truth distribution that the model is approximating, whereas the actual "ground truth" (the residual distribution) co-evolves with the model's predictions. This means that improvements in NLL conflate (i) *reducing* prediction error $\varepsilon(x)$ with (ii) *better* matching the distribution of that error. As a result, we argue that more fine-grained metrics and evaluation protocols are needed to disentangle these two aspects.

4.2 Extending to the General Case

So far, we have considered a deterministic scenario where the relationship between X and Y contains no aleatoric uncertainty and any distribution to be learned is purely the model's residual distribution. In practice, however, most real-world problems involve non-zero aleatoric noise: Y can vary even for fixed X. Therefore the full distribution must include both aleatoric and epistemic components of uncertainty.

Conceptually, one way to handle this mix is to regard *epistemic* uncertainty as a meta-distribution over the unknown parameters of the aleatoric distribution. Formally, suppose the aleatoric distribution of $Y \mid X$ is characterized by some parameter vector $\mathbf{r}_x = (r_x^{(1)}, \ldots, r_x^{(n)})$. The model then needs to produce a predictive distribution for the aleatoric parameter vector $\mathbf{r}_x$, i.e.,

$$p(y \mid X = x, \theta) = \int_{\mathbb{R}^n} p(y \mid r)\, p(r \mid \theta, X = x)\, \mathrm{d}r. \tag{8}$$

Predicting the distribution of $\mathbf{r}_x$ is the approach is taken by most methods, e.g., Bayesian Neural Networks (Blundell et al., 2015; Gawlikowski et al., 2023), which use a nondeterministic neural network to model the distribution of aleatoric parameters.

Essentially, while the value of target variable Y is no longer assumed to be deterministic given X, we can still assume that there is an ideal mapping $x \mapsto \mathbf{r}_x$ which the model is trying to approximate based on the data. Most commonly aleatoric uncertainty is modeled in simple parametric forms (e.g., Gaussian) where $\mathbf{r}$ might be a relatively small set of parameters (e.g., mean and variance). However, more expressive distributions (Bishop, 1994) may also be used.

Modeling more than one parameter does not change the fact that there is distinction to be made between more accurately predicting the parameters and accurately assessing the error made when doing so. However, the relation of aleatoric parameters to negative log-likelihood is rarely as straightforward as in Section 4.1. To illustrate the non-trivial nature of these relations, in the next section we study an analytically tractable example for the uncertainty of the variance parameter.

4.2.1 Student-t Distribution Example

Let's assume a Gaussian aleatoric uncertainty distribution with $\mathbf{r}_x = (\mu_x, \sigma_x^2)$, and suppose somehow we are given the ideal $\mu(x) = \mu_x$ function that calculates the mean, leaving the model responsible only for determining the distribution of the parameter σ_x^2. Assuming an inverse-gamma density makes it possible to analytically reason about the effects of changing the mean and variance of σ_x^2: let $\tilde{Y}$ be the compound distribution of a Gaussian with known expectation μ, and σ^2 distributed as $\Gamma^{-1}(\alpha, \beta)$. It is then known that $p(\tilde{Y})$ is a location-scale Student-t (Gelman et al., 1995).

Proposition A. *If $\tilde{Y}$ is defined as above, then* $\mathrm{Var}(\tilde{Y}) = \mathbb{E}\left(\Gamma^{-1}(\alpha, \beta)\right) = \frac{\beta}{\alpha - 1}$,

See Appendix A.1 for proof. The proposition states that, under the assumptions, changing the variance of σ^2 has no effect on $\mathrm{Var}(\tilde{Y})$, as the latter only depends on the expected value of σ^2. However, it *does not* mean that $\mathrm{Var}(\sigma^2)$ has no effect on NLL: the shape of $p(\tilde{Y} \mid X)$ still changes in $\mathrm{Var}(\sigma^2)$. To see this, we study the differential entropy H, equivalent to the ideal NLL.

Proposition B. *Let $\tilde{Y}$ be as in Proposition A. Then as* $\mathrm{Var}(\sigma^2)$ *increases, $H(\tilde{Y})$ decreases.*

See Appendix A.2 for the proof. Note that the effect is relatively small: as illustrated in Figure 1, the entropy remains nearly constant as $\mathrm{Var}(\sigma^2)$ changes. Nonetheless, Proposition B is quite counterintuitive, since we typically expect greater uncertainty to correspond to higher entropy. However, recall from Proposition A that changing $\mathrm{Var}(\sigma^2)$ does not alter the variance of $\tilde{Y}$; it only affects its shape. Therefore, the unexpected outcome is a result of the peculiarities of differential entropy combined with the Student-t distribution's shape. Proposition B serves to further emphasize that distribution-based NLL scores can be difficult to interpret as indicators of UQ.

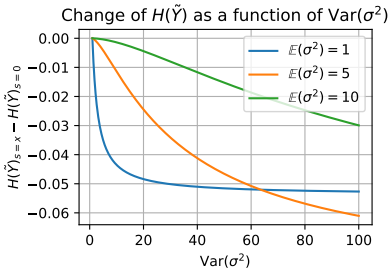

Figure 1: Change of $H(\tilde{Y})$.

Proposition A is conceptually unsurprising, as it essentially re-states the law of total variance

$$\mathrm{Var}(\tilde{Y}) = \mathbb{E}\Big[\mathrm{Var}\Big(\tilde{Y} \mid X\Big)\Big] + \mathrm{Var}\Big[\mathbb{E}\Big(\tilde{Y} \mid X\Big)\Big] \tag{9}$$

from the model's perspective, with the *uncertainty of the mean* on right-hand side assumed to be zero. However, its consequences are much more substantial than could appear at first glance. First, the result implies that given that its unbiased, the uncertainty of the predicted aleatoric variance σ^2 only affects NLL evaluation through its effect on the shape of the distribution, see Proposition B.

Second, and much more crucially, since we only have a single sample from each conditional distribution, questioning whether our prediction for σ^2 is biased only makes sense in comparison to the *unconditional*, global variance. Therefore, as long as we correctly quantify the expected error of the mean, we can use the global variance of σ^2 as an unconditional estimate for $\mathrm{Var}(\sigma^2 \mid X)$ largely without penalty, as again it only affects NLL through changing the shape of the distribution, which can affect the score either positively or negatively, as in Proposition B. This consequence is so fundamentally unexpected that we feel obligated to also verify it experimentally, see in Appendix A.3.

4.3 The Effect of Regularization

Regularization and epistemic uncertainty are closely related, both influencing model parameter selection. While UQ seeks to characterize $p(\theta \mid D)$, regularization modifies the learning process to encourage preferred parameter values. In many cases, regularization can be seen as imposing an explicit or implicit prior on model parameters. A well-known example is ridge regression, where L_2 regularization is mathematically equivalent to placing a zero-mean Gaussian prior on the model weights in a Bayesian framework.

Given this connection, it is unsuprising that several works have successfully repurposed regularization techniques for UQ. Monte Carlo Dropout (Gal & Ghahramani, 2016), for instance, leverages dropout as a means of estimating uncertainty, effectively treating it as a Bayesian approximation. Notably, BNNs themselves apply a similar weight-sampling process, though they typically rely on Gaussian rather than binary dropout-based sampling.

However, regularization is primarily designed to improve generalization and point prediction accuracy. If regularization-based UQ methods also lead to systematic reductions in prediction error, then improvements in metrics like NLL may be misleading. This reinforces our broader argument: if we cannot distinguish between gains in prediction accuracy and genuine advances in uncertainty quantification, then it remains unclear whether these models truly enhance the latter.

5 Conclusions

In this work, we argued that negative log-likelihood (NLL) alone is an incomplete measure of uncertainty quantification (UQ) in regression. While NLL encourages models to match observed distributions, it does not distinguish between improved uncertainty characterization and reduced predictive error. We showed that lower NLL can result simply from smaller residual variance, even in the absence of aleatoric noise, and that factors like aleatoric parameter uncertainty or regularization can yield misleading conclusions. These findings highlight the need to separate predictive accuracy from uncertainty representation. Without this distinction, reducing residual error can give the illusion of better UQ, even in the abscence of an actual improvement of quantification.

REFERENCES

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pp. 2078–2091. PMLR, 2023.

Christopher M Bishop. Mixture density networks. 1994.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Bai-Ni Guo, Feng Qi, Jiao-Lian Zhao, and Qiu-Ming Luo. Sharp inequalities for polygamma functions. *Mathematica Slovaca*, 65(1):103–120, 2015.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pp. 1861–1869. PMLR, 2015.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Simon Kristoffersson Lind, Ziliang Xiong, Per-Erik Forssén, and Volker Krüger. Uncertainty quantification metrics for deep regression. *Pattern Recognition Letters*, 186:91–97, 2024. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2024.09.011. URL https://www.sciencedirect.com/science/article/pii/S0167865524002733.

A.V. Lazo and P. Rathie. On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory*, 24(1):120–122, 1978. doi: 10.1109/TIT.1978.1055832.

Laurens Sluijterman, Eric Cator, and Tom Heskes. How to evaluate uncertainty estimates in machine learning for regression? *Neural Networks*, 173:106203, 2024.

A APPENDIX

A.1 STUDENT-T VARIANCE

Proposition A. *Assume a normal aleatoric uncertainty distribution with* $\mathbf{r}_x = (\mu_x, \sigma_x^2)$ *and suppose we have the true mean function* $\mu(x) = \mu_x$ *at our disposal. Further assume that our prediction for the variance* σ_x^2 *follows an inverse-gamma distribution* $\Gamma^{-1}(\alpha, \beta)$*. Then denoting the marginal distribution of the predicted variable* $\tilde{Y}_x$*,*

$$\mathrm{Var}(\tilde{Y}_x) = \mathbb{E}\left(\Gamma^{-1}(\alpha, \beta)\right) = \frac{\beta}{\alpha - 1}. \tag{10}$$

Proof. It is known (Gelman et al., 1995) that if Z is the compound distribution of a normal distribution with parameters (μ, σ^2) and $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, then

$$(Z - \mu)\left(\frac{\beta}{\alpha}\right)^{-\frac{1}{2}} \sim t(\nu = 2\alpha), \tag{11}$$

where $t(\nu)$ is Student's t-distribution with ν degrees of freedom. Further if $Z \sim t(\nu)$, then

$$\mathrm{Var}(Z) = \frac{\nu}{\nu - 2}, \tag{12}$$

and if $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$ then

$$\mathbb{E}(\sigma^2) = \frac{\beta}{\alpha - 1}. \tag{13}$$

From Equation (11), the variance of $\tilde{Y}$ is then

$$\mathrm{Var}(\tilde{Y}_x) = \frac{\nu}{\nu - 2} \cdot \frac{\beta}{\alpha} = \frac{2\alpha}{2\alpha - 2} \cdot \frac{\beta}{\alpha} = \frac{\beta}{\alpha - 1} = \mathbb{E}\left(\sigma^2\right). \tag{14}$$

□

A.2 IDEAL NLL WITH KNOWN MEAN AND ASSUMED INVERSE-GAMMA VARIANCE

First, we need to prove a bound for the half-integer difference of trigamma functions ψ_1 as a lemma.

Lemma B. *For* $z > 0$*,*

$$\frac{1}{(2z+1)(z+1)} + \frac{1}{z^2} - \frac{4}{(2z+1)^2} \geq \psi_1(z) - \psi_1\left(z + \frac{1}{2}\right), \tag{15}$$

where $\psi_1(z)$ *is the trigamma function* $\psi_1(z) = \frac{d^2}{dz^2} \log \Gamma(z)$*.*

Proof.

$$\psi_1(z) - \psi_1\left(z + \frac{1}{2}\right) = \int_z^{z+\frac{1}{2}} -\psi_2(t)\, dt\ , \tag{16}$$

where ψ_2 is the polygamma function of order 2, i.e., $\psi_2(z) = \frac{d^3}{dz^3} \log \Gamma(z)$.

It is known (Guo et al., 2015) that ψ_2 is negative on $\mathbb{R}^+$, and that

$$|\psi_2(z)| \leq \frac{1}{\left(z + \frac{1}{2}\right)^2} + \frac{2}{z^3}\ . \tag{17}$$

Therefore,

$$\int_z^{z+\frac{1}{2}} -\psi_2(t)\, dt \leq \int_z^{z+\frac{1}{2}} \frac{1}{\left(t + \frac{1}{2}\right)^2} + \frac{2}{t^3}\, \mathrm{d}t = \frac{1}{(2z+1)(z+1)} + \frac{1}{z^2} - \frac{4}{(2z+1)^2}\ . \tag{18}$$

□

Proposition C. *Assume a normal aleatoric uncertainty distribution with $\mathbf{r}_x = (\mu_x, \sigma_x^2)$ and suppose we have the true mean function $\mu(x) = \mu_x$ at our disposal. Further assume that our prediction for the variance σ_x^2 follows an inverse-gamma distribution $\Gamma^{-1}(\alpha, \beta)$. Then denoting the marginal distribution of the predicted variable $\tilde{Y}_x$, the differential entropy $H(\tilde{Y}_x)$ decreases monotonically as the variance of the distribution $\Gamma^{-1}(\alpha, \beta)$ increases.*

Proof. The differential entropy of the t distribution can be expressed (Lazo & Rathie, 1978) as

$$H(Z) = \frac{\nu+1}{2}\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right] + \log\left[\sqrt{\nu}\,\mathbf{B}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right], \tag{19}$$

where $Z \sim t(\nu)$, ψ is the digamma function $\psi(z) = \frac{d}{dz}\log\Gamma(z)$ and $\mathbf{B}$ is the Beta function $B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+Z_2)}$. Further, by the properties of differential entropy,

$$H(Z) = H((Z-d)c) - \log|c|. \tag{20}$$

Therefore, using Equation (11),

$$H\left(\tilde{Y}\right) = \left(\alpha + \frac{1}{2}\right)\left[\psi\left(\alpha + \frac{1}{2}\right) - \psi(\alpha)\right] + \log\left[\sqrt{2\alpha}\,\mathbf{B}\left(\alpha, \frac{1}{2}\right)\right] + \frac{1}{2}\log\left(\frac{\beta}{\alpha}\right). \tag{21}$$

We can reparametrize the inverse-gamma distribution using the formula for its mean $m = \frac{\beta}{\alpha-1}$ and variance $s = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ as

$$\alpha = 2 + \frac{m^2}{s} \tag{22}$$

$$\beta = m + \frac{m^3}{s}, \tag{23}$$

resulting in

$$\frac{\beta}{\alpha} = \frac{ms + m^3}{2s + m^2}. \tag{24}$$

We are interested in the behavior of the differential entropy of Equation (21) when keeping m constant and increasing s. Therefore, we further express s as a function of m and ν. Remember from Equation 11 that $\nu = 2\alpha$. Then,

$$\alpha = \frac{\nu}{2} \tag{25}$$

$$s = \frac{m^2}{\alpha - 2} = \frac{2m^2}{\nu - 4} \tag{26}$$

$$\nu = \frac{2m^2}{s} + 4 \tag{27}$$

$$\frac{\beta}{\alpha} = \frac{m\left(\frac{2m^2}{\nu-4}\right) + m^3}{2\left(\frac{2m^2}{\nu-4}\right) + m^2} = \frac{m(\nu-2)}{\nu}, \tag{28}$$

and

$$H\left(\tilde{Y}\right) = \frac{\nu+1}{2}\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right] + \log\left[\sqrt{\nu}\,\mathbf{B}\left(\frac{\nu}{2}, \frac{1}{2}\right)\right] + \frac{1}{2}\log\left(\frac{m(\nu-2)}{\nu}\right). \tag{29}$$

From Equation (27), as s increases, the value of ν decreases, converging to 4. Therefore, we expect the value of $H(\tilde{Y})$ to decrease as ν decreases. To prove this, we need to verify that sign of the derivative is positive, i.e., that:

$$\frac{d}{d\nu}H(\tilde{Y}) = \frac{\nu+1}{4}\left[\psi_1\left(\frac{\nu+1}{2}\right) - \psi_1\left(\frac{\nu}{2}\right)\right] + \frac{1}{2(\nu-2)} > 0, \tag{30}$$

where $\psi_1(z)$ is the trigamma function $\psi_1(z) = \frac{d^2}{dz^2} \log \Gamma(z)$. This is equivalent to showing that

$$\frac{2}{(\nu-2)(\nu+1)} > \psi_1\left(\frac{\nu}{2}\right) - \psi_1\left(\frac{\nu+1}{2}\right). \tag{31}$$

By Lemma B, we have

$$\frac{2}{(\nu+1)(\nu+2)} + \frac{4}{\nu^2} - \frac{4}{(\nu+1)^2} \geq \psi_1\left(\frac{\nu}{2}\right) - \psi_1\left(\frac{\nu+1}{2}\right). \tag{32}$$

It is then straightforward to check that for $\nu > 4$, we indeed have

$$\frac{2}{(\nu-2)(\nu+1)} > \frac{2}{(\nu+1)(\nu+2)} + \frac{4}{\nu^2} - \frac{4}{(\nu+1)^2}. \tag{33}$$

□

A.3 Experiment about the Uncertainty of Aleatoric Parameters

Let us define the data generating process as

$$X \sim \mathcal{U}(0, 5), \tag{34}$$
$$\mu(x) = \sin(x\pi), \tag{35}$$
$$\sigma^2(x) = \cos^2(2x\pi) + \frac{1}{2}, \tag{36}$$

and

$$p(Y \mid X = x) = \mathcal{N}(\mu(x), \sigma^2(x)). \tag{37}$$

We then numerically measure the NLL score of models of varying accuracy, represented by

$$\tilde{\mu}(x) \sim \mathcal{N}(\mu(x), s_1) \tag{38}$$

and

$$\tilde{\sigma}^2(x) = \tilde{\sigma}_e^2(x) + \tilde{\sigma}_a^2(x), \tag{39}$$

where

$$\tilde{\sigma}_e^2(x) = s_1, \text{ and} \tag{40}$$
$$\tilde{\sigma}_a^2(x) \sim \Gamma^{-1}\left(2 + \frac{\sigma^4(x)}{s_2}, \sigma^2(x) + \frac{\sigma^6(x)}{s_2}\right), \tag{41}$$

cf. Equations (22) and (24). Note that Equation 40 essentially assumes perfect knowledge about the uncertainty of μ.

We run each measurement with a sample of $n = 10000$ points. Please see the data distribution visualized on Figure 2. Figure 3 displays the NLL scores measured with different amounts of uncertainty for the aleatoric parameters μ and σ^2. As we can see, there is barely any effect of s_2 on the resulting NLL scores, especially when we are also dealing with an uncertain μ.

In Figure 4, we display scores for various rows of Figure 3, along with NLL scores of a model with an estimate for σ^2 that is not conditioned on X, rather is set to the global variance of the data. As we can see, the resulting scores are basically almost independent of whether we use conditional or global estimates for σ^2, and mostly depend on s_1, the uncertainty of μ, which we assumed perfect knowledge of.

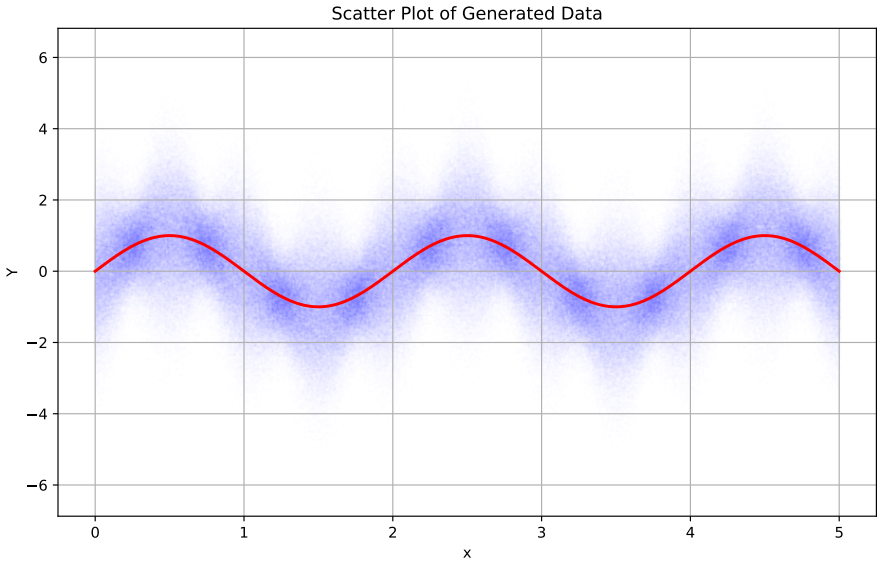

Figure 2: Distribution of the data.

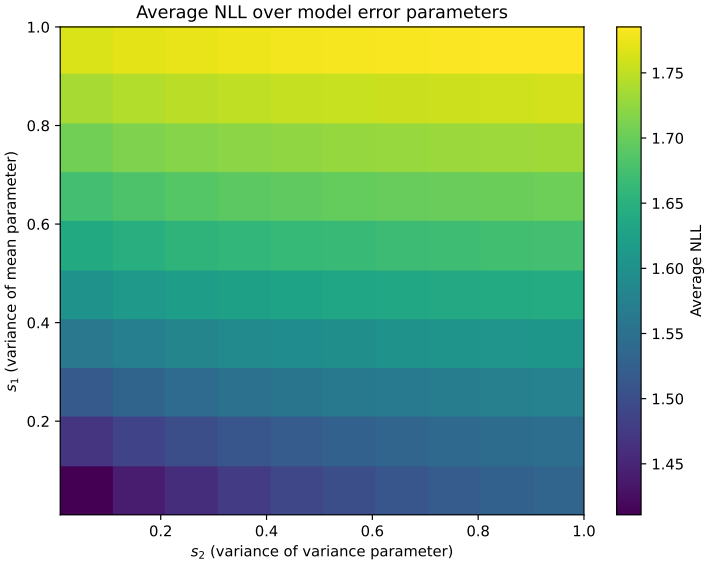

Figure 3: Visualizing NLL as a function of s_1, s_2.

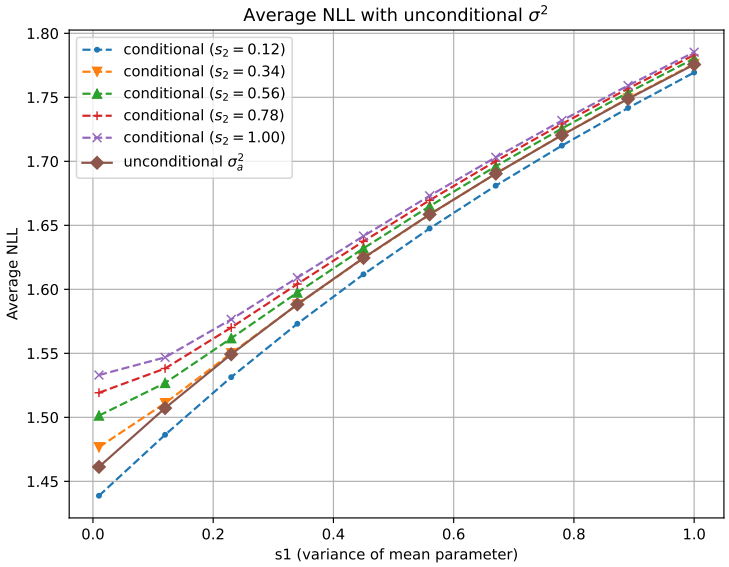

Figure 4: Visualizing NLL with an unconditional estimate for σ^2.