On Second-Order Scoring Rules for Epistemic Uncertainty Quantification

Viktor Bengs [1 2] **Eyke Hüllermeier** [1 2] **Willem Waegeman** [3]

Abstract

It is well known that accurate probabilistic predictors can be trained through empirical risk minimisation with proper scoring rules as loss functions. While such learners capture so-called aleatoric uncertainty of predictions, various machine learning methods have recently been developed with the goal to let the learner also represent its epistemic uncertainty, i.e., the uncertainty caused by a lack of knowledge and data. An emerging branch of the literature proposes the use of a second-order learner that provides predictions in terms of distributions on probability distributions. However, recent work has revealed serious theoretical shortcomings for second-order predictors based on loss minimisation. In this paper, we generalise these findings and prove a more fundamental result: There seems to be no loss function that provides an incentive for a second-order learner to faithfully represent its epistemic uncertainty in the same manner as proper scoring rules do for standard (first-order) learners. As a main mathematical tool to prove this result, we introduce the generalised notion of second-order scoring rules.

1. Introduction

The representation and quantification of uncertainty in machine learning, most notably of predictive uncertainty in the setting of supervised learning, has recently attracted increasing attention (Hüllermeier & Waegeman, 2021). Going beyond standard probabilistic prediction, various methods have been proposed that seek to distinguish between so-called aleatoric and epistemic uncertainty (Senge et al., 2014; Kendall & Gal, 2017). One way to do so is to learn second-order predictors $H : \mathcal{X} \longrightarrow \mathbb{P}(\mathbb{P}(\mathcal{Y}))$ mapping a query instance $\boldsymbol{x}$ to a probability distribution on the probability distributions over the outcome space $\mathcal{Y}$. This is motivated as follows: Assuming that outcomes cannot be predicted deterministically, and hence taking a conditional probability $p^* = p^*(\cdot \mid \boldsymbol{x})$ on $\mathcal{Y}$ as the ground-truth, it is natural to train a probabilistic predictor producing estimates $\hat{p} = \hat{p}(\cdot \mid \boldsymbol{x})$. Such estimates capture aleatoric uncertainty about the actual outcome $y \in \mathcal{Y}$, i.e., inherent randomness that the learner cannot get rid of (even with perfect knowledge about p^*, the outcome y remains random to some extent). However, it does not allow the learner to express its epistemic uncertainty, namely, its lack of knowledge about how accurately $\hat{p}$ approximates p^*. To capture this uncertainty as well, the learner is allowed to predict a second-order distribution Q. In other words, instead of committing to a single (point) prediction $\hat{p}$, the learner assigns probabilities $Q(p)$ to all candidate distributions p^*.

How to train a second-order predictor H on empirical data in the form of tuples $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, as commonly assumed in supervised learning? To this end, several authors have proposed extensions of empirical risk minimisation, i.e., to find a predictor H that minimises the (regularised) loss on the training data. Obviously, this requires a second-order loss function L_2 that compares second-order predictions with actual outcomes: $L_2(Q, y)$ is the loss suffered by the learner when predicting $Q = H(\boldsymbol{x})$ and observing outcome y. Different loss functions of this kind have been proposed for classification (Sensoy et al., 2018; Malinin & Gales, 2018; 2019; Malinin et al., 2020b; Charpentier et al., 2020; Huseljic et al., 2020; Kopetzki et al., 2021; Tsiligkaridis, 2021; Bao et al., 2021; Hammam et al., 2022) and regression (Amini et al., 2020; Ma et al., 2021; Malinin et al., 2020a; Charpentier et al., 2022; Oh & Shin, 2022; Pandey & Yu, 2022).

Focusing on the classification setting, Bengs et al. (2022) have shown theoretical shortcomings of second-order loss minimisation. In particular, they prove that the second-order loss functions proposed in the literature do not incentivise the learner to predict its epistemic uncertainty in a faithful way. Similar issues have been revealed by Meinert et al. (2022) for empirical loss minimisation in the regression setting. While criticising specific types of losses, none of these papers strictly excludes the existence of other loss functions that may provide the right incentive for the learner.

In this paper, we therefore strive for a more general result,

[1]LMU Munich, Germany [2]Munich Center for Machine Learning, Germany [3]Ghent University, Belgium. Correspondence to: Viktor Bengs <viktor.bengs@lmu.de>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

which applies to any kind of loss function and to both the classification and regression setting. To this end, we introduce second-order scoring rules as our main mathematical tool. For the case of standard (first-order) probabilistic predictions, it is well known that loss functions in the form of proper scoring rules (such as log-loss in classification and squared error loss in regression) provide exactly the right incentive to the learner: To minimise such a loss in expectation, the learner has to provide unbiased predictions of ground-truth probabilities $p^*(\cdot \mid \boldsymbol{x})$.

Transferring this notion from the aleatoric to the epistemic level, we ask the following question: Is there a second-order loss L_2 that incentivises the learner to be honest in the sense of predicting $Q = H(\boldsymbol{x})$ whenever Q corresponds to its actual belief about the ground-truth distribution $p^*(\cdot \mid \boldsymbol{x})$? Our main result is again a negative answer to this question: There seems to be no meaningful second-order loss function (scoring rule) incentivising the learner to faithfully reveal its true beliefs on the epistemic level.

It should be mentioned that Whitcomb & Benson (1996) point out a similar issue for the Continuous Ranked Probability Score (Matheson & Winkler, 1976; Gneiting & Raftery, 2007). However, they do not study the general question of existence of a second-order scoring rule of the above kind as rigorously as we do in this work.

2. Setting and Notation

We assume a standard supervised learning setting with instance (or feature) space $\mathcal{X}$, label (or outcome) space $\mathcal{Y}$, and training data $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{Y}$ can either correspond to a classification task (i.e., $\mathcal{Y} = \{y_1, \ldots, y_K\}$ for some $K \in \mathbb{N}_{\geq 2}$) or a regression task (i.e., $\mathcal{Y} = \mathbb{R}$). Following the classical setting, we also assume that the data is generated i.i.d. according to an underlying joint probability p^* on $\mathcal{X} \times \mathcal{Y}$, i.e., each $z^{(n)} = (\boldsymbol{x}^{(n)}, y^{(n)})$ is a realisation of $Z = (X, Y) \sim p^*$. Correspondingly, each instance $\boldsymbol{x} \in \mathcal{X}$ is associated with a conditional distribution $p^*(\cdot \mid \boldsymbol{x})$ on $\mathcal{Y}$, such that $p^*(y \mid \boldsymbol{x})$ is the probability to observe label y as an outcome given $\boldsymbol{x}$.

Let $\mathbb{P}(\Omega)$ denote the set of probability distributions on the measurable space $(\Omega, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra on Ω. We write $\mathbb{P}_1(\mathcal{Y}) := \mathbb{P}(\mathcal{Y})$ for the set of all probability distributions over $\mathcal{Y}$ and $\mathbb{P}_2(\mathcal{Y}) := \mathbb{P}(\mathbb{P}(\mathcal{Y}))$ for the set of all probability distributions over $\mathbb{P}(\mathcal{Y})$. For sake of convenience, we also define $\mathbb{P}_0(\mathcal{Y}) := \mathcal{Y}$. We refer to the elements in $\mathbb{P}_1(\mathcal{Y})$ as first-order distributions, while the elements in $\mathbb{P}_2(\mathcal{Y})$ are referred to as second-order distributions. We shall use lowercase letters, e.g. $\hat{p}, p$, for elements of the former, and uppercase letters, e.g. $\hat{Q}, Q$, for elements of the latter. The Dirac measure at $y \in \mathbb{P}_0(\mathcal{Y})$ is denoted by $\delta_y \in \mathbb{P}_1(\mathcal{Y})$; likewise, $\delta_p \in \mathbb{P}_2(\mathcal{Y})$ denotes the Dirac measure at $p \in \mathbb{P}_1(\mathcal{Y})$, where the underlying space of the Dirac measure should be clear from the context. Finally, we write $\overline{\mathbb{R}} = [-\infty, \infty]$ for the extended real line.

Learning Predictive First-Order Models Suppose a *hypothesis space* $\mathcal{H}_1 \subset \mathbb{P}_1(\mathcal{Y})^{\mathcal{X}} = \{h : \mathcal{X} \to \mathbb{P}_1(\mathcal{Y})\}$ to be given. Thus, a hypothesis h in $\mathcal{H}_1$ maps instances $\boldsymbol{x} \in \mathcal{X}$ to probability distributions on outcomes (first-order distributions). In standard supervised learning, the goal of the learner is to induce a hypothesis (predictive model) with low (first-order) risk

$$R_1(h) := \int_{\mathcal{X} \times \mathcal{Y}} L_1(h(\boldsymbol{x}), y) \, \mathrm{d}p^*(\boldsymbol{x}, y) \ , \tag{1}$$

where $L_1 : \mathbb{P}_1(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$ is a (first-order) loss function. The choice of a hypothesis is commonly guided by the empirical risk

$$R_{1,emp}(h) := N^{-1} \sum_{n=1}^{N} L_1\left(h(\boldsymbol{x}^{(n)}), y^{(n)}\right) \ , \tag{2}$$

i.e., the performance of a hypothesis on the training data. However, since $R_{1,emp}(h)$ is only an estimation of the true risk $R_1(h)$, the empirical risk minimiser

$$\hat{h} := \operatorname*{argmin}_{h \in \mathcal{H}_1} R_{1,emp}(h)$$

(or any other predictor) favoured by the learner will normally not coincide with the true risk minimiser (Bayes predictor)

$$h^* := \operatorname*{argmin}_{h \in \mathcal{H}_1} R_1(h).$$

Correspondingly, there remains (epistemic) uncertainty regarding h^* as well as the approximation quality of $\hat{h}$ (in the sense of its proximity to h^*) and the predictions $\hat{p}(\cdot \mid \boldsymbol{x}) = \hat{h}(\boldsymbol{x})$ produced by this hypothesis.

Example 2.1. In the classification setting with $\mathcal{Y} = \{y_1, \ldots, y_K\}$, appropriate first-order loss functions are the Brier score or the cross-entropy loss:

$$L_1^{\texttt{Brier}}(p, y) = \sum_{k=1}^{K} (p(y_k) - 1_{\{y_k = y\}})^2 \, , \tag{3}$$

$$L_1^{\texttt{CE}}(p, y) = -\sum_{k=1}^{K} 1_{\{y_k = y\}} \log(p(y_k)) \, , \tag{4}$$

where $1_{\{\cdot\}}$ is the indicator function. Both have the appealing property that the optimal hypothesis $h^*(\boldsymbol{x})$ coincides with the conditional class distribution $p^*(\cdot \mid \boldsymbol{x})$. Other suitable first-order losses include the spherical score, Winkler's score, or the Beta score. We refer to Gneiting & Raftery (2007) for an overview, who also provide examples of appropriate first-order loss functions for the case of regression (i.e., $\mathcal{Y} = \mathbb{R}$).

Learning Predictive Second-Order Models Quite recently, there has been much interest in predictive models of second-order, i.e., mappings from instances $\boldsymbol{x} \in \mathcal{X}$ to probability distributions on probability distributions over the outcomes (second-order distributions). Formally, a hypothesis space $\mathcal{H}_2 \subset \mathbb{P}_2(\mathcal{Y})^{\mathcal{X}} = \{H : \mathcal{X} \to \mathbb{P}_2(\mathcal{Y})\}$ is considered. Thus, $H(\boldsymbol{x})$ assigns a probability to each distribution $p(\cdot \,|\, \boldsymbol{x}) \in \mathbb{P}_1(\mathcal{Y})$, and the more certain the learner about the true conditional distribution $p^*(\cdot \,|\, \boldsymbol{x})$, the more concentrated or "peaked" $H(\boldsymbol{x})$ is.

In light of this, the basic idea of *direct* epistemic uncertainty prediction is to try to learn a second-order predictor in the "classical" way through loss minimisation, just like a first-order predictor. Formally, a second-order loss function

$$L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R} \tag{5}$$

is specified, which compares second-order predictions $H(\boldsymbol{x})$ with (zero-order) observations y, such that minimising L_2 on the training data $\mathcal{D}$ yields a "good" second-order predictor. Formally, the minimiser $\hat{H}$ of the empirical risk induced by L_2, i.e.,

$$R_{2,emp}(H) \coloneqq N^{-1} \sum_{n=1}^{N} L_2\left(H(\boldsymbol{x}^{(n)}), y^{(n)}\right) \quad , \tag{6}$$

over the considered hypothesis space $\mathcal{H}_2$ should provide accurate predictions for a given $\boldsymbol{x}$ by means of $\mathbb{E}_{p\sim\hat{H}(\boldsymbol{x})}\mathbb{E}_{Y\sim p}[Y]$, while reporting the second-order (epistemic) uncertainty in a reasonable and faithful manner. Preferably, these properties should be reflected by the true risk minimiser $H^* \coloneqq \operatorname{argmin}_{H\in\mathcal{H}_2} R_2(H)$, where

$$R_2(H) \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} L_2(H(\boldsymbol{x}), y)\, \mathrm{d}p^*(\boldsymbol{x}, y) \tag{7}$$

is the (second-order) risk induced by the loss L_2.

Usually, the hypotheses in $\mathcal{H}_2$ are mappings from $\mathcal{X}$ to a parametrised family of second-order distributions. More precisely, the image of these mappings is $\mathbb{P}_2(\mathcal{M})$, where $\mathcal{M}$ is some specific parameter space such that $\mathbb{P}_2(\mathcal{M})$ is in fact a strict subset of $\mathbb{P}_2(\mathcal{Y})$. Each element $Q \in \mathbb{P}_2(\mathcal{M})$ can be encoded by means of a parameter vector $\mathbf{m} \in \mathcal{M}$, i.e., $Q = Q_{\mathbf{m}}$. Thus, the hypotheses in $\mathcal{H}_2$ are encoded by a mapping from instances $\boldsymbol{x} \in \mathcal{X}$ to a parameter vector $\mathbf{m}$. In light of this, the second-order distributions $Q \in \mathbb{P}_2(\mathcal{M})$ have usually support only on a strict subset of $\mathbb{P}_1(\mathcal{Y})$ due to the parameterisation. This support is again often a parameterised family of first-order distributions $\mathbb{P}_1(\Theta)$, where Θ is yet another parameter space.

Example 2.2 (Classification). Consider the classification setting with $\mathcal{Y} = \{y_1, \ldots, y_K\}$, where the goal is to learn a predictive second-order model. Here, the most commonly used parameterised class of second-order distributions $\mathbb{P}_2(\mathcal{M})$ is the set of Dirichlet distributions with parameter space

$$\mathcal{M} = \left\{\mathbf{m} = (m_1, \ldots, m_K) \,|\, m_i > 0,\ i = 1, \ldots, K\right\}$$

having support on the (first-order) categorical distributions $\mathbb{P}_1(\Theta)$, where

$$\Theta = \left\{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in [0,1]^K \,|\, \|\boldsymbol{\theta}\|_1 = 1\right\}$$

(see (Sensoy et al., 2018; Malinin & Gales, 2018; 2019; Malinin et al., 2020b; Charpentier et al., 2020; Huseljic et al., 2020; Kopetzki et al., 2021; Tsiligkaridis, 2021; Bao et al., 2021; Hammam et al., 2022)). In this case, $\mathbb{P}_1(\mathcal{Y}) = \mathbb{P}_1(\Theta)$.

In this setting, losses of the following kind have been suggested:

$$L_2^{\mathtt{Bay}}\big(Q, y\big) = \mathbb{E}_{p\sim Q}\, L_1\big(p, y\big) + \lambda\, d_{KL}(Q, Q_0)\,, \tag{8}$$

where $\lambda \geq 0$ is some regularisation parameter, $L_1 : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ is some appropriate first-order loss, $d_{KL} : \mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_2(\mathcal{Y}) \to \mathbb{R}$ the KL-divergence, and Q_0 the uniform distribution on $\mathbb{P}_2(\mathcal{Y})$. The idea is that the first component in (8) enforces correct predictions, which, however, might favour peaked second-order distributions. Therefore, the second component in (8) acts as a countermeasure, since it penalises deviations from the most non-peaked second-order distribution, namely the uniform distribution. Other variants have also been proposed for this component, e.g. based on the Fisher information matrix (Tsiligkaridis, 2021) or the Hilbert-Schmidt independence criterion (Bao et al., 2021).

Example 2.3 (Regression). Consider the regression setting, i.e., $\mathcal{Y} = \mathbb{R}$, where again the goal is to learn a predictive second-order model. Amini et al. (2020) published the pioneering work in this regard, using normal-inverse gamma (NIG) distributions for the parameterised class of second-order distributions $\mathbb{P}_2(\mathcal{M})$, such that

$$\mathcal{M} = \big\{\mathbf{m} = (m_1, \ldots, m_4) \,|\, m_1 \in \mathbb{R}, m_2, > 0 \\ m_3 > 1, m_4 > 0 \,\big\}\,.$$

Note that Amini et al. (2020) denote (m_1, m_2, m_3, m_4) by $(\gamma, \nu, \alpha, \beta)$. Accordingly, this second-order distribution is essentially a distribution over the set of Gaussian distributions, i.e.,

$$\mathbb{P}_1(\Theta) = \{\mathrm{N}(\mu, \sigma^2) \,|\, (\mu, \sigma) \in \Theta\},$$

where $\Theta = \{(\mu, \sigma) \,|\, \mu \in \mathbb{R}, \sigma > 0\}$ is the set of (location-scale) parameters of Gaussian distributions.

The second-order loss function suggested in this regard is

$$L^{\mathtt{DER}}\big(\mathbf{m}, y\big) = L^{\mathrm{t}}\big(\mathbf{m}, y\big) + \lambda \cdot \mathrm{PEN}\big(\mathbf{m}, y\big), \tag{9}$$

where

$$L^{\mathrm{t}}(\mathbf{m}, y) = {}^1\!/_2 \log\left({}^{\pi}\!/_{m_2}\right) - m_3 \log(m_{2,4}) + (m_3 + {}^1\!/_2) \log\left((y - m_1)^2 m_2 + m_{2,4}\right) + \log\left({}^{\Gamma(m_3)}\!/_{\Gamma(m_3+\frac{1}{2})}\right), \tag{10}$$

$m_{2,4} := 2m_4(1 + m_2)$ and

$$\mathrm{PEN}(\mathbf{m}, y) = |m_1 - y| \cdot (m_3 + 2m_2).$$

Here, L^{t} is the negative log-likelihood function of a Student-t distribution with location parameter m_1, scale parameter $2m_3$ and $\frac{m_{2,4}}{2m_2 m_3}$ degrees of freedom. Similarly, as for the class of second-order loss functions in the classification setting in (8), the first component in (9) should enforce correct predictions, while the second component prevents the use of too peaked second-order distributions. For an NIG distribution, this can be achieved by penalising too large values of m_2 and m_3, as the variance terms of an NIG distribution depend reciprocally on these, respectively.

Follow-up papers on (Amini et al., 2020) adjust the loss in (9) by replacing the negative log-likelihood with squared loss (Oh & Shin, 2022), by changing the regularisation term (Pandey & Yu, 2022), or by considering a mixture of NIG distributions instead of a single one (Ma et al., 2021). Another line of research modifies (8) to the regression setting (Malinin et al., 2020a; Charpentier et al., 2022).

3. Scoring Rules for First-order Losses

In this section, we review the essential concepts of proper scoring rules, which is a class of (first-order) loss functions incentivising the learner to predict probabilities in an unbiased way. Here, unbiased means that the learner minimises expected loss if (and only if) it predicts the true (conditional) probability distribution.

Definition 3.1. A (first-order) scoring rule $S_1 : \mathbb{P}_1(\mathcal{Y}) \times \mathbb{P}_1(\mathcal{Y}) \to \overline{\mathbb{R}}$ based on the (first-order) loss $L_1 : \mathbb{P}_1(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$, such that $L_1(p, \cdot)$ is $\mathbb{P}_1(\mathcal{Y})$-quasi-integrable[1] for all $p \in \mathbb{P}_1(\mathcal{Y})$, is given for all $\hat{p}, p \in \mathbb{P}_1(\mathcal{Y})$ by

$$S_1(\hat{p}, p) = \mathbb{E}_{Y \sim p}[L_1(\hat{p}, Y)]. \tag{11}$$

The second component (i.e., p) of a scoring rule represents the target distribution or ground-truth $p^*(\cdot \mid \boldsymbol{x})$, while the first component (i.e., $\hat{p}$) represents the predicted distribution, e.g. $\hat{p}(\cdot \mid \boldsymbol{x}) = \hat{h}(\boldsymbol{x})$. Thus, integrating (11) over the distribution of the instances $\boldsymbol{x}$ leads to the (first-order) risk in (1), i.e., $R_1(\hat{h}) = \int_{\mathcal{X}} S_1(\hat{h}(\boldsymbol{x}), p(y|\boldsymbol{x}))\, \mathrm{d}p_X(\boldsymbol{x})$, where p_X denotes the distribution over the instances.

[1] A function defined on $\mathcal{Y}$ and taking values in the extended real line is $\mathbb{P}_1(\mathcal{Y})$-quasi-integrable if it is measurable w.r.t. $\mathcal{A}$ and is quasi-integrable w.r.t. all $p \in \mathbb{P}_1(\mathcal{Y})$.

Note that in the literature it is more common to refer to the loss function L_1 as the scoring rule, while S_1 is referred to as the expected score (Gneiting & Raftery, 2007; Ovcharov, 2018). However, to make the distinction between loss function and scoring rule even clearer, we will stick with the notion in Definition 3.1.

Structural properties imposed on a scoring rule allow to assess the goodness-of-fit between the distributions by means of the scoring rule.

Definition 3.2. A (first-order) scoring rule S_1 is called

- *regular* w.r.t. the class $\mathbb{P}_1(\mathcal{Y})$ if $S_1(\hat{p}, p) \in \mathbb{R}$ for all $\hat{p}, p \in \mathbb{P}_1(\mathcal{Y})$ except possibly that $S_1(\hat{p}, p) = \infty$ if $\hat{p} \neq p$.
- *proper* w.r.t. the class $\mathbb{P}_1(\mathcal{Y})$ if

$$S_1(\hat{p}, p) \geq S_1(p, p) \quad \text{for all } \hat{p}, p \in \mathbb{P}_1(\mathcal{Y}). \tag{12}$$

- *strictly proper* w.r.t. the class $\mathbb{P}_1(\mathcal{Y})$ if it is proper and

$$S_1(\hat{p}, p) > S_1(p, p) \quad \text{for all } \hat{p} \neq p. \tag{13}$$

Regular scoring rules assign finite scores, except that a prediction might receive an infinite score, e.g., if an event claimed to be impossible is realised. For proper scoring rules predicting the target distribution gives the best expectation, while strictly proper scoring rules ensure that no other prediction can achieve this value. From an uncertainty awareness perspective, the remark by Gneiting & Raftery (2007) in this regard is enlightening: "If S is proper, then the forecaster who wishes to maximize the expected score is encouraged to be honest and to volunteer his or her true beliefs."[2] Or, similarly, the one by Ovcharov (2018) regarding (strictly) proper scoring rules: "By being maximized in expectation at the true prediction, they incentivize a forecaster to truthfully report his private information."[3]

4. Scoring Rules for Second-order Losses

Inspired by the uncertainty awareness perspective of first-order scoring rules, we ask whether one can define a similar scoring rule for second-order losses $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$. Apparently, such a scoring rule needs to be a mapping from $\mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_2(\mathcal{Y})$ to $\overline{\mathbb{R}}$ to maintain the same spirit.

Definition 4.1. A (second-order) scoring rule $S_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_2(\mathcal{Y}) \to \overline{\mathbb{R}}$ based on the (second-order) loss $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$, such that $L_2(Q, \cdot)$ is $\mathbb{P}_2(\mathcal{Y})$-quasi-integrable for all $Q \in \mathbb{P}_2(\mathcal{Y})$, is given for all $\hat{Q}, Q \in \mathbb{P}_2(\mathcal{Y})$ by

$$S_2(\hat{Q}, Q) = \mathbb{E}_{p \sim Q}\left[\mathbb{E}_{Y \sim p}[L_2(\hat{Q}, Y)]\right]. \tag{14}$$

[2] Gneiting & Raftery (2007) consider the scenario of maximising the score instead of minimising it as we do in this paper, which is more in line with the standard approach in machine learning.

[3] Note that Ovcharov (2018) also considers maximising the score instead of minimising it as we do in this paper.

Note that the definition of a proper second-order scoring rule is quite different from what is known in the literature as a (local) proper scoring rule of order two (Parry et al., 2012; Ehm & Gneiting, 2012). The latter refers to (first-order) scoring rules, where the score depends on the predictive density only through its value and the values of its first as well as second derivative at the realizing event. Moreover, the second-order losses (or scoring rules) we consider have a different signature as the ones in (Murphy & von Holstein, 1971), which are of the form $\tilde{L}_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_1(\mathcal{Y}) \to \overline{\mathbb{R}}$.

Compared to (11) the definition in (14) involves an additional expectation w.r.t. the second-order distribution of the second component (i.e., Q). Again, the second component represents a target (second-order) distribution, while the first component (i.e., $\hat{Q}$) represents a predicted (second-order) distribution for a given instance $\boldsymbol{x}$, e.g. $\hat{Q}(\boldsymbol{x}) = \hat{H}(\boldsymbol{x})$. Unlike first-order distributions, however, there is no such thing as a ground-truth second-order distribution. Nevertheless, the above definition of a second-order scoring rule allows a similar close connection to the second-order risk as for first-order scoring rules: Suppose one would know a priori that the underlying conditional distributions, considered as random functions, are distributed according to a known second-order distribution $Q_{\boldsymbol{x}}$ (varying with the instances $\boldsymbol{x}$). Then, S_2 as in Definition 4.1 relates to the risk induced by L_2 (see (7)) as follows:

$$\begin{aligned}\int_{\mathcal{X}} S_2(\hat{H}(\boldsymbol{x}), Q_{\boldsymbol{x}})\,\mathrm{d}p_X(\boldsymbol{x}) \\ = \int_{\mathcal{X}} \int_{\mathbb{P}_1(\mathcal{Y})} R_2(\hat{H}(\boldsymbol{x}), p)\,\mathrm{d}Q_{\boldsymbol{x}}(p)\mathrm{d}p_X(\boldsymbol{x}),\end{aligned} \tag{15}$$

where $R_2(\hat{H}(\boldsymbol{x}), p) = \int_{\mathcal{Y}} L_2(\hat{H}(\boldsymbol{x}), y)\,\mathrm{d}p(y)$ is the conditional risk of $\hat{H}$ if $p \in \mathbb{P}_1(\mathcal{Y})$ is the ground-truth conditional distribution. Thus, a risk minimising learner is automatically encouraged to minimise the second-order scoring rule in case the target second-order distribution Q is known.

This connection is perhaps even more clarified in the case of learning without an instance space[4], where the latter equation (15) boils down to

$$S_2(\hat{H}, Q) = \int_{\mathbb{P}_1(\mathcal{Y})} R_2(\hat{H}, p)\,\mathrm{d}Q(p).$$

Akin to the first-order case (see Definition 3.2) we can specify structural properties of a second-order scoring rule.

Definition 4.2. A (second-order) scoring rule S_2 is called

- *regular* w.r.t. the class $\mathbb{P}_2(\mathcal{Y})$ if $S_2(\hat{Q}, Q) \in \mathbb{R}$ for any $\hat{Q}, Q \in \mathbb{P}_2(\mathcal{Y})$ except possibly that $S_2(\hat{Q}, Q) = \infty$ if $\hat{Q} \neq Q$.
- *proper* w.r.t. the class $\mathbb{P}_2(\mathcal{Y})$ if
$$S_2(\hat{Q}, Q) \geq S_2(Q, Q) \quad \text{for all } \hat{Q}, Q \in \mathbb{P}_2(\mathcal{Y}). \tag{16}$$
- *strictly proper* w.r.t. the class $\mathbb{P}_2(\mathcal{Y})$ if it is proper and
$$S_2(\hat{Q}, Q) > S_2(Q, Q) \quad \text{for all } \hat{Q} \neq Q. \tag{17}$$

Given the similarity of (16) (or (17)) to (12) (or (13)) as well as the similar relationship of second-order proper scoring rules to the second-order risk as in the case of first-order, we can provide a similar remark as above for proper first-order scoring rules regarding uncertainty quantification for the second-order case. That is, if S_2 is proper, then the learner that wishes to minimise the expected score is encouraged to be honest and to volunteer its true beliefs (represented by the second component). In other words, if the second-order distribution Q is the (subjective) belief of the learner, then the best score is only obtained by using $\hat{Q} = Q$ as a prediction, i.e., sticking to its own belief.

In the following, we derive a characterization of (strictly) proper second-order scoring rules similar to those known for (strictly) proper first-order scoring rules (see Theorem 1 in (Gneiting & Raftery, 2007)). To this end, we need the definition of a concave functional on $\mathbb{P}_2(\mathcal{Y})$ and its induced supertangent (or supergradient).

Definition 4.3. (i) A function $G : \mathbb{P}_2(\mathcal{Y}) \to \mathbb{R}$ is *concave* if for all $\lambda \in [0, 1], Q, \widetilde{Q} \in \mathbb{P}_2(\mathcal{Y})$ it holds that

$$G(\lambda Q + (1 - \lambda)\widetilde{Q}) \geq \lambda G(Q) + (1 - \lambda)G(\widetilde{Q}).$$

It is *strictly concave* if the latter holds with equality only in the case where $Q = \widetilde{Q}$.

(ii) A function $G^*(Q, \cdot) : \mathcal{Y} \to \overline{\mathbb{R}}$ is a *supertangent* of G at $\widetilde{Q} \in \mathbb{P}_2(\mathcal{Y})$ if it is integrable w.r.t. $\widetilde{Q}$, quasi-integrable w.r.t. to all $Q \in \mathbb{P}_2(\mathcal{Y})$ and for all $Q \in \mathbb{P}_2(\mathcal{Y})$ it holds that

$$\begin{aligned}G(Q) \leq G(\widetilde{Q}) \\ + \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G^*(\widetilde{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}(Q - \widetilde{Q})(p).\end{aligned} \tag{18}$$

In case the inequality in (18) is strict for $Q \neq \widetilde{Q}$, then G^* is called a *strict supertangent* of G.

Equipped with this we can show the following characterization of (strictly) proper second-order scoring rules, the proof of which can be found in Appendix B.1.

Theorem 4.4. *A scoring rule S_2 based on the (second-order) loss $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ is (strictly) proper iff there exists a (strictly) concave function $G_2 : \mathbb{P}_2(\mathcal{Y}) \to \mathbb{R}$ such that*

$$\begin{aligned}L_2(Q, y) = G_2(Q) + G_2^*(Q, y) \\ - \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G_2^*(Q, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p)\end{aligned} \tag{19}$$

[4]Equivalently, we may assume an instance space $\mathcal{X} = \{x_0\}$ consisting of only a single instance, which is observed over and over again (and can therefore be ignored, as it does not carry any information).

for all $Q \in \mathbb{P}_2(\mathcal{Y})$ and $y \in \mathcal{Y}$, where $G_2^*(Q,\cdot) : \mathcal{Y} \to \overline{\mathbb{R}}$ is a (strict) supertangent of G at Q.

This theorem states in essence that a second-order scoring rule S_2 induced by a second-order loss L_2 is (strictly) proper if and only if $G_2(\cdot) = S_2(\cdot,\cdot)$ is (strictly) concave and $L_2(Q,\cdot)$ is a (strict) supertangent of G_2 at Q for all $Q \in \mathbb{P}_2(\mathcal{Y})$.

Another characterization of (strictly) proper scoring rules can be derived by means of (strictly) order sensitive functions (Nau, 1985; Ovcharov, 2018).

Definition 4.5. A function $S : \mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_2(\mathcal{Y}) \to \overline{\mathbb{R}}$ is (strictly) order sensitive if the function

$$\begin{aligned} f :& [0,1] \to \overline{\mathbb{R}} \\ & \lambda \mapsto S((1-\lambda)Q' + \lambda Q, Q) \end{aligned} \tag{20}$$

is (strictly) monotonically decreasing for all second-order distributions $Q, Q' \in \mathbb{P}_2(\mathcal{Y})$.

If S is a (second-order) scoring rule, this property states that the score increases steadily as one moves away from the target distribution. Unsurprisingly, there is a close connection between (strict) propriety and (strict) order sensitivity of a scoring rule, as shown in the following theorem (for the proof see Appendix B.2).

Theorem 4.6. *A scoring rule S_2 based on the (second-order) loss $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ is (strictly) proper iff S_2 is (strictly) order sensitive.*

Finally, the following property of (strictly) proper second-order scoring rules is useful as it allows to normalise the scores if necessary.

Lemma 4.7. *If $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$ induces a (strictly) proper scoring rule S_2, then $\tilde{L}_2(Q,y) = cL_2(Q,y) + g(y)$ for any constant $c > 0$ and $\mathbb{P}_2(\mathcal{Y})$-integrable function $g : \mathcal{Y} \to \mathbb{R}$ induces a (strictly) proper scoring rule $\tilde{S}_2$.*

The proof follows directly from Definition 4.2 and is therefore omitted.

5. Non-existence of Proper Second-order Scoring Rules

In this section, we use the characterizations of (strictly) proper second-order scoring rules derived above to show negative results regarding the existence of a reasonable second-order loss L_2 such that the induced second-order scoring rule is (strictly) proper. Here, reasonable refers to loss functions that are desired from an optimization perspective, i.e., almost continuous, as well as an uncertainty penalization perspective, which we shall discuss in more detail after each theoretical result.

5.1. Classification

We start with the classification setting, i.e., $\mathcal{Y} = \{y_1, \dots, y_K\}$ for some $K \in \mathbb{N}_{\geq 2}$. Note that any probability distribution p on $\mathcal{Y}$ is characterised by a probability mass function, which we shall also denote by p.

Theorem 5.1. *There exists no loss function $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ such that the induced second-order scoring rule S_2 is proper if either of the following holds for L_2 :*

(i) $L_2(\cdot, y'')$ is almost continuous for all $y'' \in \mathcal{Y}$ and for all $y \in \mathcal{Y}$, $Q, \overline{Q} \in \mathbb{P}_2(\mathcal{Y})$, it holds that

$$L_2(Q,y) < L_2(\overline{Q},y) \tag{21}$$

iff $\mathbb{E}_{p\sim Q}[p(y)] > \mathbb{E}_{p\sim \overline{Q}}[p(y)]$.

(ii) there exist $y \in \mathcal{Y}$ and $Q, \overline{Q} \in \mathbb{P}_2(\mathcal{Y})$ such that

$$L_2(\overline{Q},y) < L_2(Q,y) \;\&\; \mathbb{E}_{p\sim\overline{Q}}[p(y)] < \mathbb{E}_{p\sim Q}[p(y)] \tag{22}$$

and

$$\begin{aligned} &\sum_{y_k \neq y} \big(L_2(\overline{Q},y_k) - L_2(\overline{Q},y)\big) \\ &\leq \sum_{y_k \neq y} \big(L_2(Q,y_k) - L_2(Q,y)\big). \end{aligned} \tag{23}$$

Proof. Case (i). Assume that L_2 satisfies (i). Note that (16) is a for-all-condition, so that if we assume (i) to hold and can find two second-order distributions such that (16) is not fulfilled, we are done. As (i) is also a for-all-condition, it is sufficient to construct two second-order distributions that essentially consider a binary classification setting, i.e., they only have support on first-order distributions which in turn have only support on two fixed classes, but violate (16).

Consequently, we restrict the second-order distributions to the binary classification case with $K = 2$, by considering the subset of $\mathbb{P}_2(\mathcal{Y})$ which has only support on first-order distributions which in turn have only support on two fixed classes. For ease of notation, let us use the encoding $y_1 = 0$ and $y_2 = 1$, so that $\mathcal{Y} = \{0,1\}$.

In light of Theorem 4.6 the function in (20) for $S = S_2$ needs to be (strictly) monotonically decreasing if S_2 is (strictly) proper. Thus, for all $\lambda \in [0,1], Q, Q' \in \mathbb{P}_2(\mathcal{Y})$ it must hold that

$$S_2(\widetilde{Q}, Q) \geq S_2(Q,Q),$$

where we abbreviated $\widetilde{Q} = \lambda Q + (1-\lambda)Q'$. This is equivalent to

$$\begin{aligned} &\int_{\mathbb{P}_1(\mathcal{Y})} L_2(\widetilde{Q},0)p(0) + L_2(\widetilde{Q},1)p(1)\,\mathrm{d}Q(p) \\ &\geq \int_{\mathbb{P}_1(\mathcal{Y})} L_2(Q,0)p(0) + L_2(Q,1)p(1)\,\mathrm{d}Q(p) \end{aligned}$$

which due to $p(0) = 1 - p(1)$ for all $p \in \mathbb{P}_1(\mathcal{Y})$ can be further rewritten to

$$\int_{\mathbb{P}_1(\mathcal{Y})} L_2(\widetilde{Q}, 0) + \Big(L_2(\widetilde{Q}, 1) - L_2(\widetilde{Q}, 0)\Big)p(1)\,\mathrm{d}Q(p) \geq \int_{\mathbb{P}_1(\mathcal{Y})} L_2(Q, 0) + \Big(L_2(Q, 1) - L_2(Q, 0)\Big)p(1)\,\mathrm{d}Q(p). \tag{24}$$

Choose $Q, Q' \in \mathbb{P}_2(\mathcal{Y})$ and λ such that

- $\widetilde{Q} = \delta_{\delta_0}$, i.e., the second-order Dirac measure, which puts all its mass on the first-order Dirac measure at 0,
- $0 < \mathbb{E}_{p\sim Q}[p(1)]$, which can be achieved as soon as Q assigns mass to first-order distributions, which are assigning positive mass to class 1,
- $\mathbb{E}_{p\sim Q}[p(1)] < \left(1 + \frac{L_2(\widetilde{Q},1) - L_2(Q,1)}{L_2(Q,0) - L_2(\widetilde{Q},0)}\right)^{-1} < 1$, which can be achieved, since $L_2(\cdot, y'')$ is by assumption almost continuous for all $y'' \in \mathcal{Y}$, and (21) together with the previous two properties of Q and $\widetilde{Q}$ implies

$$\min\Big(L_2(\widetilde{Q}, 1) - L_2(Q, 1), L_2(Q, 0) - L_2(\widetilde{Q}, 0)\Big) > 0.$$

Then, (24) is violated, since

$$\begin{aligned}
&\mathbb{E}_{p\sim Q}[p(1)] < \left(1 + \frac{L_2(\widetilde{Q}, 1) - L_2(Q, 1)}{L_2(Q, 0) - L_2(\widetilde{Q}, 0)}\right)^{-1} \\
&\Leftrightarrow L_2(\widetilde{Q}, 0) + \mathbb{E}_{p\sim Q}[p(1)]\Big(L_2(\widetilde{Q}, 1) - L_2(\widetilde{Q}, 0)\Big) \\
&< L_2(Q, 0) + \mathbb{E}_{p\sim Q}[p(1)]\Big(L_2(Q, 1) - L_2(Q, 0)\Big) \\
&\Leftrightarrow \int_{\mathbb{P}_1(\mathcal{Y})} L_2(\widetilde{Q}, 0) + \Big(L_2(\widetilde{Q}, 1) - L_2(\widetilde{Q}, 0)\Big)p(1)\,\mathrm{d}Q(p) \\
&< \int_{\mathbb{P}_1(\mathcal{Y})} L_2(Q, 0) + \Big(L_2(Q, 1) - L_2(Q, 0)\Big)p(1)\,\mathrm{d}Q(p).
\end{aligned}$$

Case (ii). Assume that L_2 satisfies (ii). Similarly to (24) we can derive that

$$\begin{aligned}
&\int_{\mathbb{P}_1(\mathcal{Y})} L_2(\widetilde{Q}, y) \\
&\quad + \sum_{y_k \neq y} \big(L_2(\widetilde{Q}, y_k) - L_2(\widetilde{Q}, y)\big)p(y_k)\,\mathrm{d}Q(p) \\
&\geq \int_{\mathbb{P}_1(\mathcal{Y})} L_2(Q, y) \\
&\quad + \sum_{y_k \neq y} \big(L_2(Q, y_k) - L_2(Q, y)\big)p(y_k)\,\mathrm{d}Q(p).
\end{aligned} \tag{25}$$

must hold if S_2 is proper, since $p(y) = 1 - \sum_{y_k \neq y} p(y_k)$. However, if $\lambda \in [0, 1], Q' \in \mathbb{P}_2(\mathcal{Y})$ are such that $\widetilde{Q} = \overline{Q}$, (which is possible due to convexity of $\mathbb{P}_2(\mathcal{Y})$), then (25) is violated, due to (22) and (23). □

Note that the two cases are not entirely exhaustive, as condition (23) is required additionally to condition (22), but requiring only the latter condition would correspond to the complementary condition of (21). Nevertheless, condition (23) essentially requires the loss function to be cost-insensitive regarding any two classes, which is in the absence of additional a priori knowledge on the data set not too restrictive. For illustration purposes, consider the binary classification case, i.e., $\mathcal{Y} = \{y_1, y_2\}$, so that (23) reads (w.l.o.g. $y = y_1$) as

$$L_2(\overline{Q}, y_2) - L_2(\overline{Q}, y_1) \leq L_2(Q, y_2) - L_2(Q, y_1).$$

Now, if (22) holds, then $\overline{Q}$ assigns fewer (prediction) probability to y_1 than Q. As a consequence, $\overline{Q}$ will assign more (prediction) probability to y_2 than Q. However, if the costs for assigning probability to the class y_2 are much higher than for y_1, i.e., the loss is cost-sensitive, then the latter left-hand side $L_2(\overline{Q}, y_2) - L_2(\overline{Q}, y_1)$ will be large, but the right-hand side $L_2(Q, y_2) - L_2(Q, y_1)$ will be small, so that the inequality is likely to be violated. If the loss is cost-insensitive, however, this will not be the case, or to be more precise, we can find Q and $\overline{Q}$ such that this is not the case (condition (ii) of Theorem 5.1 associated to (23) is an existence statement).

Condition (22) is fulfilled if the second-order loss function L_2 penalises second-order point predictions (i.e., δ_p) more drastically as for instance (second-order) predictions which are slightly deviating from point predictions. This is the case for loss functions with a regularisation term that introduces a bias towards the second-order uniform distribution (Sensoy et al., 2018; Charpentier et al., 2020; Tsiligkaridis, 2021). As a consequence, the (loss-minimising) learner has a tendency to predict more flat distributions. On the other hand, the condition in (21) enforces second-order point predictions to be more concentrated for the correct class, and avoid concentration for incorrect classes, so that the learner has a tendency to predict more peaked distributions.

The results complement those of Bengs et al. (2022) for the empirical risk minimiser for the existing second-order losses (see (8)), since losses fulfilling (22) and (23) are a generalization of the Bayesian losses with a too large regularisation parameter, while losses fulfilling (21) generalise the Bayesian losses with a too low regularisation parameter.

5.2. Regression

Next, we consider the case of regression, i.e., $\mathcal{Y} = \mathbb{R}$.

Theorem 5.2. *There exists no loss function $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ such that the induced second-order scoring rule S_2 is proper if either of the following holds for L_2 :*

(i) *$L_2(\cdot, y'')$ is almost continuous for all $y'' \in \mathcal{Y}$ and for all $y \in \mathcal{Y}$, $Q, \overline{Q} \in \mathbb{P}_2(\mathcal{Y})$, it holds that*

$$L_2(Q, y) < L_2(\overline{Q}, y) \tag{26}$$

iff $|\mathbb{E}_{p\sim Q}[\mathbb{E}(p)] - y| < |\mathbb{E}_{p\sim \overline{Q}}[\mathbb{E}(p)] - y|$.

(ii) *there exist $\mu \in \mathcal{Y}$, a first-order distribution $\tilde{p} \in \mathbb{P}_1(\mathcal{Y})$ with mean μ, a second-order distribution $\overline{Q} \in \mathbb{P}_2(\mathcal{Y})$, and some $\delta > 0$ such that for almost all [5] $y \in (\mu - \delta, \mu + \delta)$ it holds that*

$$L_2(\overline{Q}, y) < L_2(\delta_{\tilde{p}}, y) \tag{27}$$

$$\text{and} \quad \int_{\left((\mu-\delta,\mu+\delta)\right)^{\complement}} L_2(\overline{Q}, y)\, \mathrm{d}\tilde{p}(y) < \int_{\left((\mu-\delta,\mu+\delta)\right)^{\complement}} L_2(\delta_p, y)\, \mathrm{d}\tilde{p}(y), \tag{28}$$

where $\left((\mu - \delta, \mu + \delta)\right)^{\complement} = \mathcal{Y} \setminus (\mu - \delta, \mu + \delta)$.

Proof. Case (i). Let $p_l, p_r \in \mathbb{P}_1(\mathcal{Y})$ be first-order distributions and $\mu^* \in \mathbb{R}$ such that

- p_l has only support on $(-\infty, \mu^*)$ and p_r has only support on (μ^*, ∞),
- p_l has expected value $\mu_l < \mu^*$ and p_r has an expected value $\mu_r > \mu^*$,
- it holds for all $\lambda \in (0, 1)$ that

$$\mathbb{E}_{Y\sim p_l}[L_2(Q_\lambda, Y) - L_2(\delta_{p_l}, Y)] > 0,$$

where $Q_\lambda = (1 - \lambda)\delta_{p_l} + \lambda\delta_{p_r}$.

This is possible, since

$$\begin{aligned}
&\mathbb{E}_{Y\sim p_l}[L_2(Q_\lambda, Y) - L_2(\delta_{p_l}, Y)] \\
&\quad = \int_{(-\infty,\mu_l]} L_2(Q_\lambda, y) - L_2(\delta_{p_l}, y)\, \mathrm{d}p_l(y) \\
&\qquad + \int_{[\mu_l,\mu^*)} L_2(Q_\lambda, y) - L_2(\delta_{p_l}, y)\, \mathrm{d}p_l(y)
\end{aligned}$$

and the first term is always positive by (26), as $\mathbb{E}_{p\sim\delta_{p_l}}[\mathbb{E}(p)] = \mu_l$ and $\mathbb{E}_{p\sim Q_\lambda}[\mathbb{E}(p)] = (1-\lambda)\mu_l + \lambda\mu_r > \mu_l$. Thus, by suitable choice of μ^*, p_l and p_r (as well as μ_l and μ_r), the second term can be designed such that it is smaller than the first in absolute terms, since $L_2(\cdot, y'')$ is by assumption almost continuous for all $y'' \in \mathcal{Y}$.

Let $\widetilde{Q} = \delta_{p_l}$, and choose $Q \in \mathbb{P}_2(\mathcal{Y})$ such that $Q = (1 - \lambda^*)\delta_{p_l} + \lambda^*\delta_{p_r}$, where

$$0 < \lambda^* < \left(1 + \frac{\mathbb{E}_{Y\sim p_r}(L_2(\widetilde{Q}, Y) - L_2(Q, Y))}{\mathbb{E}_{Y\sim p_l}(L_2(Q, Y) - L_2(\widetilde{Q}, Y))}\right)^{-1} < 1. \tag{29}$$

[5] A condition holds for *almost all* x in some set X, if the subset on which the condition does not hold has probability mass 0.

This choice of λ^* can be achieved as $\mathbb{E}_{Y\sim p_l}(L_2(Q, Y) - L_2(\widetilde{Q}, Y)) > 0$ by choice of μ^*, p_l and p_r, and (26) implies that $\mathbb{E}_{Y\sim p_r}(L_2(\widetilde{Q}, Y) - L_2(Q, Y)) > 0$, since for all y in the support of p_r it holds that $L_2(\widetilde{Q}, y) - L_2(Q, y) > 0$. Thus,

$$\begin{aligned}
&S_2(\widetilde{Q}, Q) \\
&= \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} L_2(\widetilde{Q}, y)\, \mathrm{d}p(y)\, \mathrm{d}Q(p) \\
&= (1 - \lambda^*) \int_{\mathcal{Y}} L_2(\widetilde{Q}, y)\, \mathrm{d}p_l(y) + \lambda^* \int_{\mathcal{Y}} L_2(\widetilde{Q}, y)\, \mathrm{d}p_r(y) \\
&= (1 - \lambda^*)\mathbb{E}_{Y\sim p_l}(L_2(\widetilde{Q}, Y)) + \lambda^*\mathbb{E}_{Y\sim p_r}(L_2(\widetilde{Q}, Y)) \\
&< (1 - \lambda^*)\mathbb{E}_{Y\sim p_l}(L_2(Q, Y)) + \lambda^*\mathbb{E}_{Y\sim p_r}(L_2(Q, Y)) \\
&= S_2(Q, Q),
\end{aligned}$$

where the inequality is due to (29). The latter shows that we can find two second-order distributions that violate (16).

Case (ii). Let us abbreviate $(\mu-\delta, \mu+\delta)$ by $(\mu\pm\delta)$. Choose Q to be $\delta_{\tilde{p}}$, then (27) and (28) imply that

$$\begin{aligned}
&S_2(\overline{Q}, Q) \\
&\quad = \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} L_2(\overline{Q}, y)\, \mathrm{d}p(y)\, \mathrm{d}Q(p) \\
&\quad = \int_{\mathcal{Y}} L_2(\overline{Q}, y)\, \mathrm{d}\tilde{p}(y) \\
&\quad = \int_{(\mu\pm\delta)} L_2(\overline{Q}, y)\, \mathrm{d}\tilde{p}(y) + \int_{(\mu\pm\delta)^{\complement}} L_2(\overline{Q}, y)\, \mathrm{d}\tilde{p}(y) \\
&\quad < \int_{(\mu\pm\delta)} L_2(\delta_{\tilde{p}}, y)\, \mathrm{d}\tilde{p}(y) + \int_{(\mu\pm\delta)^{\complement}} L_2(\delta_{\tilde{p}}, y)\, \mathrm{d}\tilde{p}(y) \\
&\quad = \int_{(\mu\pm\delta)} L_2(Q, y)\, \mathrm{d}\tilde{p}(y) + \int_{(\mu\pm\delta)^{\complement}} L_2(Q, y)\, \mathrm{d}\tilde{p}(y) \\
&\quad = \int_{\mathcal{Y}} L_2(Q, y)\, \mathrm{d}\tilde{p}(y) \\
&\quad = \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} L_2(Q, y)\, \mathrm{d}p(y)\, \mathrm{d}Q(p) = S_2(Q, Q).
\end{aligned}$$

Thus, S_2 is not proper. □

The two cases in Theorem 5.2 are quite similar to the ones in Theorem 5.1: the first corresponds to second-order losses that incentivise the learner to predict more flat distributions, while the second incentivises predictions of peaked distributions. The second-order loss function for the regression case suggested by Amini et al. (2020) (see Example 2.3) fulfils the conditions in the second case of Theorem 5.2 as the following proposition shows.

Proposition 5.3. *The deep evidential regression loss function $L_2^{DER} : \mathbb{P}_2(\mathcal{M}) \times \mathcal{Y} \to \mathbb{R}$ in (9) fulfils the conditions in Theorem 5.2 (ii).*

Proof. Let $\mu \in \mathbb{R}$ be arbitrary but fixed and $\sigma > 0$ be some small value. Note that an NIG distribution with parameters $m_1 = \mu$, $m_2 = \infty$ and m_3, m_4 such that $m_4/(m_3 - 1) = \sigma^2$ and m_3 sufficiently large corresponds to $\delta_{\tilde{p}}$ with $\mathbb{E}(\tilde{p}) = \mu$ and $\mathbb{V}(\tilde{p}) = \sigma^2$. However, by using for $\overline{Q}$ an NIG distribution with $\tilde{m}_1 = m_1$, $\tilde{m}_2 = 1$, $\tilde{m}_3 = 1$ and $\tilde{m}_4 = m_4$, the deep evidential regression loss of $\delta_{\tilde{p}}$ is larger for all y except for $y = \mu$ than the deep evidential regression loss of $\overline{Q}$. □

6. Conclusion

Our results confirm concerns raised by recent work regarding the conceptual meaningfulness of direct epistemic uncertainty quantification through empirical risk minimisation of second-order distributions. More precisely, unlike for the case of empirical risk minimisation of strictly proper first-order loss functions to report first-order (aleatoric) uncertainty in a faithful manner, there seems to be no strictly proper second-order loss function counterpart to report second-order (epistemic) uncertainty in a faithful manner. Although our results do not completely exclude existence, they do so for a large class of second-order loss functions, including those losses previously proposed in the literature.

The most likely explanation for the revealed existence issue is the discrepancy between the orders that these second-order loss functions exhibit: a second-order prediction $H(\boldsymbol{x})$ is evaluated in light of a zero-order observation y, skipping the intervening first-order. Thus, to make the loss function meaningful, one would rather need observations of realizations of the first-order distribution, i.e., a sample in the form of probabilities and assess the second-order prediction in light of these first-order observations. However, such data cannot exist even in principle, because the ground-truth conditional distribution is supposedly constant.

This suggests that (probabilistic) learning on the epistemic level cannot be frequentist in nature, unlike learning about the ground-truth conditional distribution on the first-order (aleatoric level). Instead, it appears that learning on the second-order (epistemic level) is necessarily Bayesian and requires a prior, which then of course has an influence on the degree of (epistemic) uncertainty.

Acknowledgements

Willem Wageman received funding from the Flemish Government under the "Onderzoeksprogramma Artificielë Intelligentie (AI) Vlaanderen" Programme.

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. In *Proc. NeurIPS, 33rd Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937, 2020.

Bao, W., Yu, Q., and Kong, Y. Evidential deep learning for open set action recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13329–13338, 2021.

Bengs, V., Hüllermeier, E., and Waegeman, W. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Proc. NeurIPS, 35th Advances in Neural Information Processing Systems*, 2022.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In *Proc. NeurIPS, 33rd Neural Information Processing Systems*, volume 33, pp. 1356–1367, 2020.

Charpentier, B., Borchert, O., Zugner, D., Geisler, S., and Günnemann, S. Natural posterior network: Deep Bayesian predictive uncertainty for exponential family distributions. In *Proc. ICLR, 10th International Conference on Learning Representations*, 2022.

Ehm, W. and Gneiting, T. Local proper scoring rules of order two. *The Annals of Statistics*, 40(1):609–637, 2012.

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

Hammam, A., Bonarens, F., Ghobadi, S. E., and Stiller, C. Predictive uncertainty quantification of deep neural networks using Dirichlet distributions. In *Computer Science in Cars Symposium*, pp. 1–10, 2022.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3.

Huseljic, D., Sick, B., Herde, M., and Kottke, D. Separation of aleatoric and epistemic uncertainty in deterministic deep neural networks. In *Proc. ICPR, 25th International Conference on Pattern Recognition*, pp. 9172–9179. IEEE, 2020.

Kendall, A. and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proc. NIPS, 30th Advances in Neural Information Processing Systems*, pp. 5574–5584, 2017.

Kopetzki, A., Charpentier, B., Zügner, D., Giri, S., and Günnemann, S. Evaluating robustness of predictive uncertainty estimation: Are Dirichlet-based models reliable? In *Proc. ICML, 38th International Conference on Machine Learning*, pp. 5707–5718, 2021.

Ma, H., Han, Z., Zhang, C., Fu, H., Zhou, J. T., and Hu, Q. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. In *Proc. NeurIPS, 34th Advances in Neural Information Processing Systems*, pp. 6881–6893, 2021.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Proc. NeurIPS, 31st Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

Malinin, A. and Gales, M. Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Proc. NeurIPS, 32nd Advances in Neural Information Processing Systems*, pp. 14520–14531, 2019.

Malinin, A., Chervontsev, S., Provilkov, I., and Gales, M. J. F. Regression prior networks. *CoRR*, abs/2006.11590, 2020a. URL https://arxiv.org/abs/2006.11590.

Malinin, A., Mlodozeniec, B., and Gales, M. Ensemble distribution distillation. In *Proc. ICLR, 8th International Conference on Learning Representations*, 2020b.

Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management Science*, 22 (10):1087–1096, 1976.

Meinert, N., Gawlikowski, J., and Lavin, A. The unreasonable effectiveness of deep evidential regression. *arXiv preprint arXiv:2205.10060*, 2022.

Murphy, A. H. and von Holstein, C.-A. S. S. Second-order probabilities and strictly proper scoring rules. *Journal of Applied Meteorology (1962-1982)*, pp. 177–179, 1971.

Nau, R. F. Should scoring rules be 'effective'? *Management Science*, 31(5):527–535, 1985.

Oh, D. and Shin, B. Improving evidential deep learning via multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7895–7903, 2022.

Ovcharov, E. Y. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1):53–79, 2018.

Pandey, D. S. and Yu, Q. Evidential conditional neural processes. *arXiv preprint arXiv:2212.00131*, 2022. URL https://arxiv.org/abs/2212.00131.

Parry, M., Dawid, A. P., and Lauritzen, S. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

Senge, R., Bösner, S., Dembczynski, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., and Hüllermeier, E. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Proc. NeurIPS, 31st Conference on Neural Information Processing Systems*, pp. 3183–3193, 2018.

Tsiligkaridis, T. Information aware max-norm Dirichlet networks for predictive uncertainty estimation. *Neural Networks*, 135:105–114, 2021.

Whitcomb, K. M. and Benson, P. G. Evaluating second-order probability judgments with strictly proper scoring rules. *Theory and Decision*, 41:165–178, 1996.

A. List of Symbols

The following table contains a list of symbols that are frequently used in the main paper as well as in the following supplementary material.

General Learning Setting	
$\mathcal{X}$	instance space
$\mathcal{Y}$	label space, either $\{y_1, \ldots, y_K\}$ for some $K \in \mathbb{N}_{\geq 2}$ for classification or $\mathcal{Y} = \mathbb{R}$ for regression
$\mathcal{D}$	training data $\left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$
p^*	data generating probability
$p^*(\cdot \mid \boldsymbol{x})$	conditional distribution or density on $\mathcal{Y}$, i.e., $p^*(y \mid \boldsymbol{x})$ probability to observe y given $\boldsymbol{x}$ in classification density of y given $\boldsymbol{x}$ in regression
$\mathbb{P}(\mathcal{Y}), \mathbb{P}_1(\mathcal{Y})$	the set of probability distributions on $\mathcal{Y}$
$\mathbb{P}_1(\Theta)$	a parameterised subset of $\mathbb{P}_1(\mathcal{Y})$ with Θ being the parameter space
First-order Learning Setting	
$\mathcal{H}_1$	(first-order) hypothesis space consisting of hypothesis $h : \mathcal{X} \longrightarrow \mathbb{P}_1(\mathcal{Y})$
L_1	loss function for first-order hypothesis, i.e., $L_1 : \mathbb{P}_1(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$
$R_{1,emp}(\cdot)$	empirical risk of a first-order hypothesis (cf. (2))
$R_1(\cdot)$	risk or expected loss of a first-order hypothesis (cf. (1))
$\hat{h}$	empirical risk minimiser, i.e., $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}_1} R_{1,emp}(h)$
h^*	true risk minimiser or Bayes predictor, i.e., $h^* = \operatorname{argmin}_{h \in \mathcal{H}_1} R_1(h)$
$S_1(\cdot, \cdot)$	first-order scoring rule induced by some first-order loss function L_1, i.e., $S_1 : \mathbb{P}_1(\mathcal{Y}) \times \mathbb{P}_1(\mathcal{Y}) \to \overline{\mathbb{R}}$ with $S_1(\hat{p}, p) = \mathbb{E}_{Y \sim p}[L_1(\hat{p}, Y)]$ (see (11))
Second-order Learning Setting	
$\mathbb{P}_2(\mathcal{Y})$	the set of distributions on $\mathbb{P}_1(\mathcal{Y})$ (the set of second-order distributions)
$\mathbb{P}_2(\mathcal{M})$	a parameterised subset of $\mathbb{P}_2(\mathcal{Y})$ with $\mathcal{M}$ being the parameter space
$\mathcal{H}_2$	(second-order) hypothesis space consisting of hypothesis $H : \mathcal{X} \longrightarrow \mathbb{P}_2(\mathcal{Y})$
$Q, Q', \bar{Q}, \widetilde{Q}$	probability distributions on $\mathbb{P}_1(\mathcal{Y})$ i.e., elements of $\mathbb{P}_2(\mathcal{Y})$
Q_0	uniform distribution on $\mathbb{P}_1(\mathcal{Y})$ (an element of $\mathbb{P}_2(\mathcal{Y})$)
L_2	loss function for second-order hypothesis, i.e., $L_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$
$L_2^{\texttt{Bay}}$	Bayesian loss functions for classification setting (see (8))
$L^{\texttt{DER}}$	deep evidential regression loss functions for regression setting (see (9))
$R_{2,emp}(\cdot)$	empirical risk of a second-order hypothesis (cf. (6))
$R_2(\cdot)$	risk or expected loss of a second-order hypothesis (cf. (7))
$\hat{H}$	second-order empirical risk minimiser, i.e., $\hat{H} = \operatorname{argmin}_{H \in \mathcal{H}_2} R_{2,emp}(H)$
$S_2(\cdot, \cdot)$	second-order scoring rule induced by some second-order loss function L_2, i.e., $S_2 : \mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_2(\mathcal{Y}) \to \overline{\mathbb{R}}$ with $S_2(\hat{Q}, Q) = \mathbb{E}_{p \sim Q}\left[\mathbb{E}_{Y \sim p}[L_2(\hat{Q}, Y)]\right]$ (see (14))
Distributions & Expectations	
$\mathrm{N}(\mu, \sigma^2)$	Gaussian distribution with location parameter μ and scale parameter $\sigma > 0$
δ_y	Dirac measure at $y \in \mathcal{Y}$ (i.e., δ_y is an element $\mathbb{P}_1(\mathcal{Y})$)
δ_p	Dirac measure at $p \in \mathbb{P}_1(\mathcal{Y})$ (i.e., δ_p is an element $\mathbb{P}_2(\mathcal{Y})$)
$\mathbb{E}(p)$	expected value of the distribution $p \in \mathbb{P}_1(\mathbb{R})$, i.e., $\mathbb{E}(p) = \int y \, \mathrm{d}p(y)$
$\mathbb{V}(p)$	variance of a distribution $p \in \mathbb{P}_1(\mathbb{R})$, i.e., $\mathbb{V}(p) = \mathbb{E}\left[(p - \mathbb{E}[p])^2\right]$
$\mathbb{E}_{p \sim Q}[p(y)]$	expected probability assigned to class $y \in \mathcal{Y}$ (i.e., classification setting) according to $Q \in \mathbb{P}_2(\mathcal{Y})$, i.e., $\mathbb{E}_{p \sim Q}[p(y)] = \int_{\mathbb{P}_1(\mathcal{Y})} p(y) \mathrm{d}Q(p)$
$\mathbb{E}_{p \sim Q}[\mathbb{E}(p)]$	expected value of the expected distribution (for regression, i.e., $\mathcal{Y} = \mathbb{R}$) according to $Q \in \mathbb{P}_2(\mathcal{Y})$, i.e., $\mathbb{E}_{p \sim Q}[\mathbb{E}(p)] = \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} y \, \mathrm{d}p(y) \mathrm{d}Q(p)$
Miscellaneous	
$d_{KL}(\cdot, \cdot)$	Kullback-Leibler divergence (on $\mathbb{P}_2(\mathcal{Y}) \times \mathbb{P}_2(\mathcal{Y})$)
$L_1^{\texttt{Brier}}$	Brier score (see (3))
$L_1^{\texttt{CE}}$	Cross-entropy loss (see (4))
$L^{\texttt{t}}$	negative log-likelihood of Student-t distribution (see (10))
PEN	penalization function in deep evidential regression loss function (see below (10))

B. Missing Proofs of Section 4

B.1. Proof of Theorem 4.4

We will use the following lemma for the proof of Theorem 4.4, which we shall prove at the end of this subsection.

Lemma B.1. *If a function $G : \mathbb{P}_2(\mathcal{Y}) \to \mathbb{R}$ has a supertangent $G^*(Q, \cdot)$ at any $Q \in \mathbb{P}_2(\mathcal{Y})$, then G is concave. If the supertangent property in (18) holds with strict inequality for all $Q \neq \widetilde{Q}$, then G is strictly concave.*

Proof. Suppose S_2 (or rather L_2) fulfils the representation in (19), then

$$\begin{aligned}
&S_2(\hat{Q}, Q)\\
&\quad= \mathbb{E}_{p\sim Q}\left[\mathbb{E}_{Y\sim p}[L_2(\hat{Q}, Y)]\right]\\
&\quad= \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} L_2(\hat{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p)\\
&\quad\overset{(19)}{=} G_2(\hat{Q}) + \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} G_2^*(\hat{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p)\\
&\qquad- \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} G_2^*(\hat{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}\hat{Q}(p)\\
&\quad= G_2(\hat{Q}) + \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} G^*(\hat{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}(Q-\hat{Q})(p)\\
&\quad\overset{(18)}{\geq} G_2(Q)\\
&\quad= G_2(Q) + \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} G_2^*(Q, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p)\\
&\qquad- \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} G_2^*(Q, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p)\\
&\quad\overset{(19)}{=} \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} L_2(Q, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p)\\
&\quad= \mathbb{E}_{p\sim Q}\left[\mathbb{E}_{Y\sim p}[L_2(Q, Y)]\right]\\
&\quad= S_2(Q, Q),
\end{aligned}$$

where we used for the inequality that G_2^* is a supertangent of G at $\hat{Q}$, i.e., that (18) holds for Q. Note that the inequality is strict if G_2 is strictly concave and $\hat{Q} \neq Q$.

Conversely, suppose S_2 to be (strictly) proper scoring rule. Define G_2 by $G_2(Q) = S_2(Q, Q)$, then $G_2^*(Q, y) = L_2(Q, y)$ is a supertangent of G_2 at any $\widetilde{Q} \in \mathbb{P}_2(\mathcal{Y})$. Indeed, let $Q \in \mathbb{P}_2(\mathcal{Y})$, then

$$\begin{aligned}
&G_2(Q)\\
&\quad= S_2(Q, Q)\\
&\quad= S_2(\widetilde{Q}, \widetilde{Q}) - S_2(\widetilde{Q}, \widetilde{Q}) + S_2(\widetilde{Q}, Q) - S_2(\widetilde{Q}, Q)\\
&\qquad+ S_2(Q, Q)\\
&\quad\leq S_2(\widetilde{Q}, \widetilde{Q}) - S_2(\widetilde{Q}, \widetilde{Q}) + S_2(\widetilde{Q}, Q)\\
&\quad= S_2(\widetilde{Q}, \widetilde{Q}) + \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} L_2(\widetilde{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}(Q-\widetilde{Q})(p)\\
&\quad= G_2(\widetilde{Q}) + \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} G_2^*(\widetilde{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}(Q-\widetilde{Q})(p),
\end{aligned}$$

where for the inequality we used that S_2 is proper. This inequality is strict for all $\widetilde{Q} \neq Q$ if S_2 is strictly proper. Thus, G_2 is (strictly) concave due to Lemma B.1. The definitions of G_2 and G_2^* directly imply that L_2 has the representation in (19), since $G_2(Q) = \int_{\mathbb{P}_1(\mathcal{Y})}\int_{\mathcal{Y}} G_2^*(Q, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p)$.

□

Proof of Lemma B.1. Let $Q, \widetilde{Q} \in \mathbb{P}_2(\mathcal{Y})$ and $\lambda \in [0,1]$ be arbitrary but fixed. Abbreviate $Q_\lambda = (1-\lambda)Q + \lambda\widetilde{Q}$. Then, since G^* is by assumption a supertangent of G for any element of $\mathbb{P}_2(\mathcal{Y})$, it holds by the supertangent property (see (18)) that

$$G(Q) \leq G(Q_\lambda) + \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G^*(Q_\lambda, y)\,\mathrm{d}p(y)\,\mathrm{d}(Q - Q_\lambda)(p),$$
$$G(\widetilde{Q}) \leq G(Q_\lambda) + \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G^*(Q_\lambda, y)\,\mathrm{d}p(y)\,\mathrm{d}(\widetilde{Q} - Q_\lambda)(p).$$

If we consider the convex combination of these two inequalities and noting that $Q - Q_\lambda = \lambda Q - \lambda\widetilde{Q}$ as well as $\widetilde{Q} - Q_\lambda = (1-\lambda)\widetilde{Q} - (1-\lambda)Q$, we obtain

$$\begin{aligned}
(1-\lambda)G(Q) + \lambda G(\widetilde{Q}) \leq\ & (1-\lambda)G(Q_\lambda) && + (1-\lambda)\int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G^*(Q_\lambda, y)\,\mathrm{d}p(y)\,\mathrm{d}(Q - Q_\lambda)(p) \\
& + \lambda G(Q_\lambda) && + \lambda\int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G^*(Q_\lambda, y)\,\mathrm{d}p(y)\,\mathrm{d}(\widetilde{Q} - Q_\lambda)(p) \\
=\ & (1-\lambda)G(Q_\lambda) && + (1-\lambda)\lambda\int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G^*(Q_\lambda, y)\,\mathrm{d}p(y)\,\mathrm{d}(Q - \widetilde{Q})(p) \\
& + \lambda G(Q_\lambda) && + (1-\lambda)\lambda\int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} G^*(Q_\lambda, y)\,\mathrm{d}p(y)\,\mathrm{d}(\underbrace{\widetilde{Q} - Q}_{=-(Q-\widetilde{Q})})(p) \\
=\ & G(Q_\lambda).
\end{aligned}$$

Thus, G is concave according to Definition 4.3.

If G^* is a strict supertangent, then all inequalities above are strict if $Q \neq \widetilde{Q}$ and $\lambda \in (0,1)$ and consequently G is strictly concave in this case. □

B.2. Proof of Theorem 4.6

Proof. Suppose S_2 is order sensitive. Since $\mathbb{P}_2(\mathcal{Y})$ is convex, we can represent any element $\hat{Q}$ by a suitable convex combination of a target second-order distribution Q and another suitable second-order distribution $\widetilde{Q}$. Formally, for all $\hat{Q}, Q$ there exist $\lambda \in [0,1]$ and $\widetilde{Q}$ such that $\hat{Q} = (1-\lambda)\widetilde{Q} + \lambda Q$. Thus,

$$S_2(\hat{Q}, Q) = S_2((1-\lambda)\widetilde{Q} + \lambda Q, Q) \geq S_2(Q, Q),$$

which implies that S_2 is proper. This inequality is strict if S_2 is strictly order sensitive and $\hat{Q} \neq Q$ implying that S_2 is strictly proper.

Now, assume that S_2 is proper. Note that any scoring-rule S_2 is (convex) linear in its second argument: For all $\lambda \in [0,1]$ and $Q, Q', \hat{Q} \in \mathbb{P}_2(\mathcal{Y})$ it holds that

$$\begin{aligned}
& S_2(\hat{Q}, \lambda Q + (1-\lambda)Q') \\
&\quad = \mathbb{E}_{p\sim\lambda Q+(1-\lambda)Q'}\left[\mathbb{E}_{Y\sim p}[L_2(\hat{Q}, Y)]\right] \\
&\quad = \int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} L_2(\hat{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}(\lambda Q + (1-\lambda)Q')(p) \\
&\quad = \lambda\int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} L_2(\hat{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}Q(p) \\
&\qquad + (1-\lambda)\int_{\mathbb{P}_1(\mathcal{Y})} \int_{\mathcal{Y}} L_2(\hat{Q}, y)\,\mathrm{d}p(y)\,\mathrm{d}Q'(p) \\
&\quad = \lambda S_2(\hat{Q}, Q) + (1-\lambda)S_2(\hat{Q}, Q').
\end{aligned}$$

Order sensitivity of a scoring rule holds if for all $\lambda \in [0,1]$ and $Q, Q' \in \mathbb{P}_2(\mathcal{Y})$

$$S_2(Q', Q) - S_2((1-\lambda)Q' + \lambda Q, Q) \geq 0. \tag{30}$$

This follows by propriety of S_2 : Abbreviate $\widetilde{Q} = (1-\lambda)Q' + \lambda Q$ and note that by convex linearity in the second argument

$$(1-\lambda)S_2(Q', Q') + \lambda S_2(Q', Q) = S_2(Q', \widetilde{Q}) \geq S_2(\widetilde{Q}, \widetilde{Q}) = (1-\lambda)S_2(\widetilde{Q}, Q') + \lambda S_2(\widetilde{Q}, Q),$$

which is equivalent to

$$S_2(Q', Q) - S_2((1-\lambda)Q' + \lambda Q, Q) \geq \frac{(1-\lambda)}{\lambda}\Big(S_2(\widetilde{Q}, Q') - S_2(Q', Q')\Big).$$

The right-hand side is non-negative since S_2 is proper, which implies (30). Finally, if S_2 is strictly proper, then (30) holds with strict inequality implying strict order sensitivity of S_2. □