Submitted to the Annals of Statistics

LOCAL PROPER SCORING RULES OF ORDER TWO

By Werner Ehm[*] and Tilmann Gneiting[*]

Institute for Frontier Areas of Psychology and Mental Health and University of Heidelberg

Scoring rules assess the quality of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes. A scoring rule is proper if it encourages truthful reporting. It is local of order k if the score depends on the predictive density only through its value and the values of its derivatives of order up to k at the realizing event. Complementing fundamental recent work by Parry, Dawid and Lauritzen (2011), we characterize the local proper scoring rules of order two relative to a broad class of Lebesgue densities on the real line, using a different approach. In a data example, we use local and non-local proper scoring rules to assess statistically postprocessed ensemble weather forecasts.

1. Introduction.

One of the major purposes of statistical analysis is to make forecasts for the future, and to provide suitable measures of the uncertainty associated with them. Consequently, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities and events (Dawid 1984). Scoring rules provide summary measures for the evaluation of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes. We take scoring rules to be positively oriented rewards that a forecaster wishes to maximize. Specifically, if the forecaster quotes the predictive distribution P and the event x materializes, her reward is $\mathrm{S}(P, x)$. The function $\mathrm{S}(P, \cdot\,)$ takes values in the extended real line, $\overline{\mathbb{R}} = [-\infty, \infty]$, and we write $\mathrm{S}(P, Q)$ for the expected value of $\mathrm{S}(P, \cdot\,)$ under Q. Suppose, then, that the forecaster's best judgement is the predictive distribution Q.

[*] The authors are grateful to Philip Dawid, Steffen Lauritzen and Matthew Parry for helpful discussions, and sharing manuscripts and presentations (cf. Remark 3.4), to Peter J. Huber for alerting us to the counterexample in Remark 3.8, and to Chris Fraley for providing R code used in Section 6. Tilmann Gneiting acknowledges support by the Alfried Krupp von und zu Behlen Foundation, and by the United States National Science Foundation under Awards ATM-0724721 and DMS-0706745, and thanks the Institute for Frontier Areas of Psychology and Mental Health in Freiburg, Germany for hospitality and travel support.

AMS 2000 subject classifications: Primary 62C99; secondary 62M20, 86A10

Keywords and phrases: density forecast, Euler equation, Fisher information, Hyvärinen score, proper scoring rule, tangent construction

The forecaster has no incentive to predict any $P \neq Q$, and is encouraged to quote her true belief, $P = Q$, if $\mathrm{S}(Q,Q) \geq \mathrm{S}(P,Q)$. A scoring rule with this property is said to be proper (Gneiting and Raftery 2007).

Our paper is concerned with local proper scoring rules for probabilistic forecasts of a real-valued quantity. Briefly, if the predictive distribution is absolutely continuous, it can be argued that $\mathrm{S}(P,x)$ ought to depend only on the behavior of the predictive density, p, in an infinitesimal neighborhood of the observation that materializes, x. Any such scoring rule is said to be local, with the logarithmic scoring rule,

$$\mathrm{S}(P,x) = \log p(x), \tag{1}$$

being the most prominent example (Good 1952). Another example is the Hyvärinen (2005) score,

$$\begin{aligned} \mathrm{S}(P,x) &= \left(\frac{p'(x)}{p(x)}\right)^2 - 2\,\frac{p''(x)}{p(x)} \\ &= -\left((\log p)'(x)\right)^2 - 2\,(\log p)''(x), \end{aligned} \tag{2}$$

which is local of order two, in the sense that it depends on the predictive density only by its value, and the values of its first and second derivative, at the observation. Similarly, the logarithmic score can be considered to be local of order zero; in fact, it is the only such score that is proper, up to equivalence (Bernardo 1979). The Hyvärinen score is also proper (Dawid and Lauritzen 2005), thus raising the question for a characterization of the local proper scoring rules of order $k \leq 2$.

In a far-reaching recent paper, Parry, Dawid and Lauritzen (2011) achieved a characterization of the key local score functions of any order $k \geq 0$. They derive these scores from the Euler-Lagrange equation of the calculus of variations, thereby obtaining natural candidates for local proper scoring rules, the actual propriety of which can be checked by additional criteria. We complement these results — for comments, see Remark 3.4 and Section 7 — by developing an alternative approach, restricting ourselves to the practically most relevant case of the local proper scoring rules of order $k \leq 2$. Our main contributions are the following: We build on a characterization of proper scoring rules via convex functionals and their (sub-)gradients, which yields the general form of the second-order local proper scoring rules in a natural tangent construction; and we specify suitable classes of scoring rules and predictive densities that allow for a full-fledged, rigorous characterization.

The remainder of the paper is organized as follows. Section 2 introduces the notions of propriety and locality in full detail. Section 3 presents our main result, in that we characterize the class of the local scoring rules of order two that are proper relative to a comprehensive family of Lebesgue densities, which includes many of the classical location-scale families on the real line. The proof is given in Section 4. Section 5 provides supplements and examples, and a data example on ensemble weather forecasts is given in Section 6. The relations to and distinctions from the work of Parry et al. (2011) are revisited in the concluding Section 7, along with a discussion of open problems and hints at possible future developments and applications.

2. Local proper scoring rules.

Initially, we consider predictive distributions on a general sample space, Ω. Let $\mathcal{A}$ be a σ-algebra of subsets of Ω, and let $\mathcal{M}$ be a class of probability measures on $(\Omega, \mathcal{A})$. A function on Ω is $\mathcal{M}$-*quasiintegrable* if it is measurable with respect to $\mathcal{A}$ and quasiintegrable with respect to all $P \in \mathcal{M}$ (Bauer 2001, p. 64). A *probabilistic forecast* or a *predictive distribution* is any probability measure $P \in \mathcal{M}$. A *scoring rule* is any extended real-valued function $\mathrm{S} : \mathcal{M} \times \Omega \to \overline{\mathbb{R}}$ such that $\mathrm{S}(P, \cdot)$ is $\mathcal{M}$-quasiintegrable for all $P \in \mathcal{M}$. Hence, if the predictive distribution is P and the event ω materializes, the forecaster's reward is $\mathrm{S}(P, \omega)$. We define

$$\mathrm{S}(P, Q) = \int \mathrm{S}(P, \omega) \, dQ(\omega)$$

as the expected score under Q when the probabilistic forecast is P. This is a well defined extended real-valued quantity, because $\mathrm{S}(P, \cdot)$ is quasi-integrable with respect to Q.

Definition 2.1. The scoring rule S is *proper* relative to $\mathcal{M}$ if

$$\mathrm{S}(Q, Q) \geq \mathrm{S}(P, Q) \quad \text{for all} \quad P, Q \in \mathcal{M}.$$

It is *strictly proper* relative to $\mathcal{M}$ if $\mathrm{S}(Q, Q) \geq \mathrm{S}(P, Q)$ with equality if and only if $P = Q$.

The term proper was coined by Winkler and Murphy (1968), while the general idea can be traced to Brier (1950) and Good (1952). Dawid (2008) provides a concise history of proper scoring rules, which includes major contributions by the subjective school of probability as well as meteorologists.

A scoring rule can be thought of as *local* if $\mathrm{S}(P, \omega)$ depends on the predictive distribution, P, only through its behavior in an infinitesimal neighborhood of the verifying observation, ω. Bernardo (1979, p. 689) argued in this vein, noting that "when assessing the worthiness of a scientist's final conclusions, only the probability he attaches to a small interval containing the true value should be taken into account." In the context of predictive densities, the class $\mathcal{M}$ is a family of probability measures that are absolutely continuous with respect to a σ-finite measure μ on $(\Omega, \mathcal{A})$. We then identify a probabilistic forecast $P \in \mathcal{M}$ with its μ-density, p, which we call a *predictive density* or a *density forecast*. The classical example of a local proper scoring rule is the aforementioned logarithmic score, which can be interpreted as a predictive likelihood, and is strictly proper relative to any such class $\mathcal{M}$.

Hereinafter, we restrict attention to the case in which the sample space Ω is the real line, $\mathcal{A}$ is the Borel σ-algebra, μ is the Lebesgue measure, and $\mathcal{M}$ corresponds to some class of Borel probability measures that admit a unique, smooth Lebesgue density, p. Accordingly, we will consider $\mathcal{M}$ as a class of densities rather than measures, and we may write $\mathrm{S}(p, \cdot)$. The logarithmic score (1) and the Hyvärinen score (2) admit particularly simple analytic forms in terms of the log likelihood, $\log p(x)$, and its derivatives, which are fundamental objects of statistical inference. Therefore, we define locality in terms of these quantities.

Definition 2.2. Let k be a nonnegative integer, and let $\mathcal{M}$ be a class of probability densities with respect to the Lebesgue measure on $\mathbb{R}$ that are everywhere strictly positive and admit derivatives up to order k. A scoring rule S for the class $\mathcal{M}$ then is *local* of *order* k if there exists a function $\mathrm{s} : \mathbb{R}^{2+k} \to \overline{\mathbb{R}}$, which we call a *scoring function*, such that

$$\mathrm{S}(p, x) = \mathrm{s}(x, \log p(x), \ldots, (\log p)^{(k)}(x))$$

for every $p \in \mathcal{M}$ and $x \in \mathbb{R}$.

An alternative notion of locality, which allows the predictive density, p, to have zeroes, would take the arguments of the scoring function as $x, p(x), \ldots, p^{(k)}(x)$. However, in addition to being natural and facilitating the technicalitites, the assumption of strict positivity avoids pathologies, as will be seen in Remark 3.8 below.

As propriety can only be assessed relative to a specified class of predictive densities, we now introduce a suitable family.

Definition 2.3. Let $\mathcal{P}$ denote the class of all probability densities, p, with respect to the Lebesgue measure on $\mathbb{R}$ that satisfy the following conditions:

(P1) p is strictly positive on $\mathbb{R}$;
(P2) p admits four continuous derivatives on $\mathbb{R}$;

(P3) for every $m > 0$ and $j = 0, 1, \ldots, 4$,

$$\lim_{x \to \pm\infty} |x|^m p^{(j)}(x) = 0;$$

(P4) there exists a constant $a = a(p) > 0$ such that

$$\lim_{x \to \pm\infty} |x|^{-a} \frac{p^{(j)}(x)}{p(x)} = 0 \quad \text{for} \quad j = 1, \ldots, 4.$$

The class $\mathcal{P}$ is quite broad and includes many well known densities, such as all normal and logistic densities, the corresponding skew variants (Genton 2004), and finite mixtures of these densities. In particular, $\mathcal{P}$ is convex; see Section 4.1.

In the following, we do not systematically distinguish a scoring rule, S, and the corresponding scoring function, s, both of which will simply be referred to as scores.

3. Characterization of the local proper scores of order two. Along with the logarithmic and the Hyvärinen score, any convex combination thereof is a local proper score of order two. However, the class of the local proper scoring rules of order two on the real line, $\mathbb{R}$, has a much richer structure, and allows for a characterization in terms of convex functionals.

3.1. *Main results.* We first introduce classes of functions that satisfy suitable polynomial growth conditions.

DEFINITION 3.1. Let k be a nonnegative integer. The class $\mathcal{R}_k$ consists of all kernels $K : \mathbb{R}^{2+k} \to \mathbb{R}$ that admit continuous partial derivatives up to order $2k$, and for which there exist finite positive constants C and r such that, whenever W stands for K or any of its partial derivatives up to order $2k$, then

$$|W(x, y_0, \ldots, y_k)| \leq C \left\{ (1 + |x|)(1 + |y_0|) \cdots (1 + |y_k|) \right\}^r$$

for all $(x, y_0, \ldots, y_k) \in \mathbb{R}^{2+k}$.

Note that the growth conditions on the functions in the class $\mathcal{R}_k$, as well as the decay conditions on the densities in the class $\mathcal{P}$ of Definition 2.3, apply to each member individually. They are not required to hold uniformly.

For a kernel $K \in \mathcal{R}_k$ and a density $q \in \mathcal{P}$, let

$$\Phi_K(q) = \int_{\mathbb{R}} K(x, \log q(x), (\log q)'(x), \ldots, (\log q)^{(k)}(x)) \, q(x) \, dx.$$

The integral exists and is finite by virtue of the growth and decay conditions imposed on K and q, respectively. Thus, any kernel $K \in \mathcal{R}_k$ induces a well-defined functional $\Phi_K : \mathcal{P} \to \mathbb{R}$. The properties of such kernel type functionals play a key role in our subsequent characterization. In stating it, we use standard abbreviations to denote the partial derivatives of a function of the form $g = g(x, y_0, \ldots, y_k)$; for example, we write $\partial_j g = \partial g / \partial y_j$ and $\partial^2_{xj} g = \partial^2 g / (\partial x \partial y_j)$. The proof is given in Section 4.

The subsequent two results are closely connected to the work of Parry et al. (2011); see Remark 3.4.

Theorem 3.2. *Let $\mathcal{P}$ denote the class of probability densities introduced in Definition 2.3.*

(a) Consider a kernel K of the form

$$K(x, y_0, y_1) = c y_0 + K_0(x, y_1), \tag{3}$$

where c is a real constant and K_0 is a real function on $\mathbb{R}^2$. If $K \in \mathcal{R}_1$ and the functional Φ_K is convex, the function $\mathrm{s} : \mathbb{R}^4 \to \mathbb{R}$, defined by

$$\mathrm{s}(x, y_0, y_1, y_2) = c y_0 + (1 - y_1 \partial_1 - \partial^2_{x1} - y_2 \partial^2_{11}) K_0(x, y_1), \tag{4}$$

represents a local score of order two that is proper relative to $\mathcal{P}$.

(b) Conversely, if $\mathrm{s} \in \mathcal{R}_2$ represents a local score of order two that is proper relative to $\mathcal{P}$, there exists a kernel $K \in \mathcal{R}_1$ of the form (3), where c is a real constant and K_0 is a real function on $\mathbb{R}^2$, such that the functional Φ_K is convex and s admits the representation (4).

(c) The above statements remain valid with convex replaced by strictly convex, and proper replaced by strictly proper.

The following sufficient condition for the functional Φ_K to be convex will be proved in Section 5.1.

Proposition 3.3. *Suppose that K is a kernel of the form (3) such that (i) $K \in \mathcal{R}_1$, (ii) $c \geq 0$, and (iii) the map $y_1 \mapsto K_0(x, y_1)$ is convex for every $x \in \mathbb{R}$. Then the functional $\Phi_K : \mathcal{P} \to \mathbb{R}$ is convex. The statement continues to hold if convex is replaced by strictly convex.*

The criterion provides a straightforward method of constructing local proper scores of order two via the basic relationship (4). For example, the kernel $K(x, y_0, y_1) = y_0$ yields the scoring function, $\mathrm{s}(x, y_0, y_1, y_2) = y_0$, that represents the logarithmic score (1). The associated functional

$$\Phi(q) = \mathrm{S}(q, q) = \int_{\mathbb{R}} \log q(x)\, q(x)\, dx$$

is the (negative of) the Shannon entropy, and the associated divergence,

$$d_{\mathrm{KL}}(p,q) = \mathrm{S}(q,q) - \mathrm{S}(p,q) = \int \log \frac{q(x)}{p(x)}\, q(x)\, dx,$$

is the Kullback-Leibler divergence. Similarly, the kernel $K(x, y_0, y_1) = y_1^2$ yields the scoring function, $\mathrm{s}(x, y_0, y_1, y_2) = -y_1^2 - 2y_2$, that represents the Hyvärinen score (2). The associated functional and divergence,

$$\int_{\mathbb{R}} \left(\frac{q'(x)}{q(x)}\right)^2 q(x)\, dx \quad \text{and} \quad d_{\mathrm{FI}}(p,q) = \int \left(\frac{p'(x)}{p(x)} - \frac{q'(x)}{q(x)}\right)^2 q(x)\, dx,$$

are the Fisher information and the Fisher information divergence (DasGupta 2008, p. 26), respectively. For further examples, see Section 5.3.

3.2. *Remarks.* It has to be emphasized that the present work owes a great deal to interactions with Philip Dawid, Steffen Lauritzen and Matthew Parry, which began with their kindly pointing out an error in our previous work (Ehm and Gneiting 2009, Addendum 2010).

Remark 3.4 (acknowledgment of priority). In the compact notation explained in Section 4, the representation (4) of the second-order local proper scoring rules can be written as

$$\mathrm{s} \;=\; K - \left[z_1 + \frac{d}{dx}\right] \partial_1 K. \tag{5}$$

We learned about this representation in a personal communication (Dawid, Parry and Lauritzen 2009). A different, though essentially equivalent expression is given in Parry et al. (2011) as the special case $t = 1$ of the general form (39) of the key local scoring rules. Concerning Proposition 3.3, compare Theorem 9.1 of Parry et al. (2011).

As noted, the original contributions of our approach will be discussed in Section 7. Here we continue with comments relating to the choice of the class $\mathcal{P}$ and the complementary roles of the kernel K as a function and a functional, thereby touching on the generality of Theorem 3.2 and Proposition 3.3.

Remark 3.5. There is a slight asymmetry in Theorem 3.2, in that under the conditions of the sufficiency part (a) the scoring function is continuous only, whereas the necessity part (b) requires it to be four times continuously differentiable. Other than this, the theorem accomplishes a full characterization of the local proper scoring rules of order two relative to the class $\mathcal{P}$ of Definition 2.3.

Remark 3.6. Part (a) of Theorem 3.2 expresses a local proper score of order two, s, in terms of a kernel, K, with suitable properties. Similarly, part (b) admits a constructive extension that finds and expresses a suitable kernel, K, in terms of a local proper score of order two, s. See Section 4.3 for the explicit construction and Example 5.2 for an illustration.

Remark 3.7. Theorem 3.2 has been stated for the special class $\mathcal{P}$ of Definition 2.3. Propriety relative to such a broad class is a fairly demanding requirement, and from this perspective, part (a) is a strong result. In contrast, part (b) would be stronger if propriety was required relative to a subclass $\mathcal{P}_0 \subset \mathcal{P}$ only. On the other hand, $\mathcal{P}_0$ must not be too narrow. An inspection of Section 4 shows that part (b) remains valid relative to any convex subclass $\mathcal{P}_0 \subset \mathcal{P}$ with the following two additional properties:

(P5) if a continuous function f on $\mathbb{R}$ with at most polynomial growth at $\pm\infty$ satisfies $\int_{\mathbb{R}} f(x)\,(p(x) - q(x))\,dx \leq 0$ for all $p, q \in \mathcal{P}_0$, then f is constant;

(P6) the richness properties of Lemma 4.11 hold for $\mathcal{P}_0$.

Condition (P5) is needed in Section 4.2, while property (P6) is required in Section 4.4. The full class $\mathcal{P}$ does satisfy these conditions.

Remark 3.8. The sufficieny part of Theorem 3.2 would be stronger if the statement applied relative to larger classes $\mathcal{P}_1 \supset \mathcal{P}$. The following adaptation of an example of Huber (1974) shows that any such extension may entail inexpected effects for strict propriety, with undesirable consequences in applications. Suppose that $\mathcal{P}$ is augmented to a convex class $\mathcal{P}_1$ that includes the densities

$$p_\alpha(x) = \begin{cases} \alpha\, g(x) & \text{if} \quad x \geq 0, \\ (1-\alpha)\, g(-x) & \text{if} \quad x < 0, \end{cases}$$

where $\alpha \in (0,1)$ and $g(x) = x^5 e^{-x}/\Gamma(6)$ for $x \geq 0$. The p_α satisfy all conditions for the class $\mathcal{P}$ except for property (P1), since $p_\alpha(x) = 0$ at $x = 0$. As the logarithmic derivatives, $p'_\alpha(x)/p_\alpha(x)$, do not depend on α, the Fisher information of p_α does not depend on α either, hence is not strictly convex as a functional on $\mathcal{P}_1$. Accordingly, the Fisher information divergence does not distinguish the p_α, $d_{\mathrm{FI}}(p_\alpha, p_\beta) = 0$ for α, $\beta \in (0,1)$, and the Hyvärinen score (2) fails to be strictly proper relative to the augmented class $\mathcal{P}_1$. In particular, strict convexity of the function $K_0(x, y_1)$ of Proposition 3.3 in y_1 does not imply strict convexity of the associated functional, unless we restrict the class of densities under consideration.

Remark 3.9. By Proposition 3.3, convexity of a kernel K of the form (3) in y_1 implies convexity of the associated functional Φ_K on the class $\mathcal{P}$. Conversely, what are the consequences of convexity of the functional Φ_K on the kernel K? The example of the logarithmic score (1) demonstrates that matters are not straightforward; here the functional Φ_K is strictly convex, yet the kernel $K(x, y_0, y_1) = y_0$ is not.

Now consider any kernel K of the form (3) for which the associated functional Φ_K is convex on $\mathcal{P}$. Do we necessarily have $c \geq 0$ then? This is indeed true if $K_0(x, y_1) = y_1^2$ represents the Hyvärinen score (2). Then by propriety

$$0 \leq \mathrm{S}(q, q) - \mathrm{S}(p, q) = c\, d_{\mathrm{KL}}(p, q) + d_{\mathrm{FI}}(p, q)$$

for all $p, q \in \mathcal{P}$, so that $c \geq 0$ is necessary if the ratio $r = d_{\mathrm{FI}}(p, q)/d_{\mathrm{KL}}(p, q)$ can attain arbitrarily small values. However, $\mathcal{P}$ contains all normal densities, and if p and q are normal with mean zero and standard deviations σ and τ, then

$$d_{\mathrm{KL}}(p, q) = \frac{\tau^2}{\sigma^2} - 1 - \log \frac{\tau^2}{\sigma^2} \qquad \text{and} \qquad d_{\mathrm{FI}}(p, q) = \frac{(1 - \tau^2/\sigma^2)^2}{\tau^2},$$

whence r can attain any positive value. The argument clearly depends on the class $\mathcal{P}$ being inclusive; it fails if $\mathcal{P}$ is replaced by a narrower class $\mathcal{P}_0$ for which the ratio r is bounded away from zero. Such is in fact possible due to a logarithmic Sobolev inequality, which asserts that for certain classes $\mathcal{P}_0 \subset \mathcal{P}$ one has $d_{\mathrm{KL}}(p, q) \leq C\, d_{\mathrm{FI}}(p, q)$ for $p, q \in \mathcal{P}_0$ with a constant C that depends only on $\mathcal{P}_0$. A corresponding reference is Villani (2009): put $u = \sqrt{q/p}$ and $d\nu(x) = p(x)\, dx$ in eq. (21.3) and consider Remark 21.4.

4. Proof of Theorem 3.2. Our point of departure is Theorem 1 of Gneiting and Raftery (2007), which can be traced to Hendrickson and Buehler (1971) and characterizes proper scoring rules by means of the subgradients of convex functionals on convex classes of probability measures. We state it in the special case where that class corresponds to the set $\mathcal{P}$ of Lebesgue densities introduced in Definition 2.3. *Throughout this section propriety is understood as propriety relative to* $\mathcal{P}$.

Theorem 4.1. *Let Φ be a convex real-valued functional on $\mathcal{P}$ with subgradient $\Phi^*(q, \cdot) : \mathbb{R} \to \mathbb{R}$ at $q \in \mathcal{P}$, that is,*

$$\Phi(p) - \Phi(q) - \int_{\mathbb{R}} \Phi^*(q, x)\, (\, p(x) - q(x))\, dx \geq 0 \quad \textit{for} \quad p, q \in \mathcal{P}.$$

Then the scoring rule

$$\mathrm{S}(q, \cdot) = \Phi^*(q, \cdot) - \int_{\mathbb{R}} \Phi^*(q, x) q(x)\, dx + \Phi(q) \tag{6}$$

is proper, and

$$\mathrm{S}(q,q) = \int_{\mathbb{R}} \mathrm{S}(q,x)\, q(x)\, dx = \Phi(q) \quad \textit{for} \quad q \in \mathcal{P}.$$

Conversely, if S *is proper then* $\Phi(q) = \mathrm{S}(q,q)$ *is a convex functional on* $\mathcal{P}$ *with subgradient* $\Phi^*(q,\cdot) = \mathrm{S}(q,\cdot)$ *at* $q \in \mathcal{P}$*, whence* S *is of the form (6). Furthermore, the above continues to hold with convex replaced by strictly convex, and proper replaced by strictly proper.*

For sufficiently regular local proper scoring rules we can compute the gradient of the corresponding functionals. Specifically, a function $G(q,\cdot) : \mathbb{R} \to \mathbb{R}$ is a *weak gradient*, or simply a *gradient*, of the functional Φ at $q \in \mathcal{P}$ if for every $p \in \mathcal{P}$

$$\frac{d}{dt}\left[\Phi(p_t)\right]\bigg|_{t=0} = \int_{\mathbb{R}} G(q,x)\,(p(x) - q(x))\, dx$$

where

$$p_t = (1-t)\,q + t\,p \qquad \text{for} \qquad t \in [0,1]. \tag{7}$$

Theorem 4.1 along with such gradient calculations gives us a construction method for local proper scores that readily elucidates their particular form. We refer to this approach as the *tangent construction* and give details in the following section, before completing the proof of Theorem 3.2 in a series of subsequent steps.

4.1. *Tangent construction of proper scores.* In what follows we use compactified notation whenever possible. As noted, we do not systematically distinguish scoring rules, S, and the corresponding scoring functions, s, both of which are referred to as scores. Log-likelihoods and their derivatives are denoted $z_0(q,x) = \log q(x)$ and

$$z_j(q,x) = (\log q)^{(j)}(x) \qquad \text{for} \qquad j = 1, 2, \ldots,$$

or simply z_0 and z_j if the density q is fixed. Clearly then,

$$z_j' = z_{j+1} = z_0^{(j+1)} \qquad \text{for} \qquad j = 0, 1, 2, \ldots,$$

where the prime denotes differentiation with respect to x. We usually suppress the differential, dx, in integrals over $x \in \mathbb{R}$, and in the corresponding integrands we omit all or part of the arguments whenever these are clear from the context. For example, given $K \in \mathcal{R}_k$ and $p, q \in \mathcal{P}$ we may abbreviate

$$\int_{\mathbb{R}} K(x, \log p(x), \ldots, (\log p)^{(k)}(x))\, q(x)\, dx$$

as

$$\int Kq \qquad (K = K_p),$$

where, evidently, $K_p = K_p(x) = K(x, \log p(x), \ldots, (\log p)^{(k)}(x))$.

Convex functionals require a convex domain. Beyond the convexity of $\mathcal{P}$, we need certain uniform estimates provided in the following lemma.

Lemma 4.2. *For every $k = 1, 2, \ldots$ there exists a polynomial $M = M(y_1, \ldots, y_k)$ of degree k such that for all $p, q \in \mathcal{P}$ and $\alpha \in [0, 1]$, the density $r_\alpha = \alpha\, p + (1 - \alpha)\, q$ satisfies*

$$\left|(\log r_\alpha)^{(k)}(x)\right| \le M\left(\max\left\{\frac{|p'(x)|}{p(x)}, \frac{|q'(x)|}{q(x)}\right\}, \ldots, \max\left\{\frac{|p^{(k)}(x)|}{p(x)}, \frac{|q^{(k)}(x)|}{q(x)}\right\}\right)$$

pointwise in $x \in \mathbb{R}$.

Proof. Let the polynomial $L(y_1, \ldots, y_k)$ of degree k be such that the k-th logarithmic derivative of a smooth function $g > 0$ can be written as

$$(\log g)^{(k)} = L\left(\frac{g'}{g}, \ldots, \frac{g^{(k)}}{g}\right),$$

where here and in the following we suppress the argument $x \in \mathbb{R}$. Define the polynomial M as L with all coefficients replaced by their absolute values. Evidently then,

$$\left|(\log r_\alpha)^{(k)}\right| \le M\left(\frac{|r_\alpha'|}{r_\alpha}, \ldots, \frac{|r_\alpha^{(k)}|}{r_\alpha}\right),$$

and it suffices to show that

$$\frac{|r_\alpha^{(j)}|}{r_\alpha} \le \max\left\{\frac{|p^{(j)}|}{p}, \frac{|q^{(j)}|}{q}\right\} \quad \text{for} \quad j = 1, \ldots, k.$$

Consider the function $f(\alpha) = (\alpha c_1 + (1 - \alpha) c_0)/(\alpha d_1 + (1 - \alpha) d_0)$, where $c_0, c_1 \in \mathbb{R}$ and $d_0, d_1 > 0$ are constants. Then

$$|f(\alpha)| \le \max\{|f(0)|, |f(1)|\} \quad \text{for} \quad \alpha \in [0, 1],$$

because $f'(\alpha) = (c_1 d_0 - c_0 d_1)/(\alpha d_1 + (1 - \alpha) d_0)^2$ does not change sign. The desired inequality follows on setting $c_0 = q^{(j)}(x)$, $c_1 = p^{(j)}(x)$, $d_0 = q(x)$ and $d_1 = p(x)$. □

Corollary 4.3. *The class $\mathcal{P}$ is convex.*

We now develop the tangent construction. The first step consists in calculating the gradients of (not necessarily convex) functionals of kernel type.

Lemma 4.4. *Let $K \in \mathcal{R}_2$. Then the gradient G of the associated functional $\Phi_K : \mathcal{P} \to \mathbb{R}$ exists at any $q \in \mathcal{P}$, and is given by*

$$G = K + \partial_0 K - \frac{1}{q}\frac{d}{dx}[q\partial_1 K] + \frac{1}{q}\frac{d^2}{dx^2}[q\partial_2 K] \quad (G = G_q,\, K = K_q). \tag{8}$$

Recall that according to our notational conventions eq. (8) means that the relation holds whenever the functions G, K and $\partial_j K$ are evaluated at arguments

$$(x, z_0, z_1, z_2) = (x, \log q(x), (\log q)'(x), (\log q)''(x)),$$

where $q \in \mathcal{P}$ and $x \in \mathbb{R}$.

Proof. Let $q \in \mathcal{P}$ be fixed. In calculating the gradient of $\Phi = \Phi_K$ at q we initially ignore all technicalities, that is, we assume that integrals are well-defined and finite, that the order of integration and differentiation can be interchanged, and that boundary terms in partial integrations vanish. Then

$$\frac{d}{dt}[\Phi(p_t)] = \int \frac{d}{dt}[K_t p_t] = \int K_t(p-q) + \int \left[\frac{d}{dt}K_t\right] p_t, \tag{9}$$

where p_t denotes the mixture density (7) and

$$K_t = K_{p_t} = K(x, \log p_t(x), (\log p_t)'(x), (\log p_t)''(x)).$$

Since $\frac{d}{dt}\log p_t = (p-q)/p_t$, the mixed derivative with respect to t and x (of order j) is given by

$$\frac{d}{dt}\left[(\log p_t)^{(j)}\right] = \left(\frac{p-q}{p_t}\right)^{(j)}.$$

The second term on the right-hand side of eq. (9) can then be computed using partial integration, in that

$$
\begin{aligned}
(10)\quad \int \left[\frac{d}{dt}K_t\right] p_t
&= \int \left[(\partial_0 K_t)\left(\frac{p-q}{p_t}\right) + (\partial_1 K_t)\left(\frac{p-q}{p_t}\right)' + (\partial_2 K_t)\left(\frac{p-q}{p_t}\right)''\right] p_t \\
&= \int (\partial_0 K_t)(p-q) - \int \left(\frac{d}{dx}[p_t \partial_1 K_t]\right)\left(\frac{p-q}{p_t}\right) \\
&\qquad + \int \left(\frac{d^2}{dx^2}[p_t \partial_2 K_t]\right)\left(\frac{p-q}{p_t}\right) \\
&= \int \left[\partial_0 K_t - \frac{1}{p_t}\frac{d}{dx}[p_t \partial_1 K_t] + \frac{1}{p_t}\frac{d^2}{dx^2}[p_t \partial_2 K_t]\right](p-q).
\end{aligned}
$$

Evaluating at $t = 0$, and noting that $p_0 = q$ and $K_0 = K_q = K$, eqns. (9) and (10) yield

$$\frac{d}{dt}\left[\Phi(p_t)\right]\bigg|_{t=0} = \int \left[K + \partial_0 K - \frac{1}{q}\frac{d}{dx}[q \partial_1 K] + \frac{1}{q}\frac{d^2}{dx^2}[q \partial_2 K]\right](p-q),$$

so that the gradient of Φ at q is indeed given by (8).

It remains to settle the technicalities. Generally, if a family $\{h(x,t) : x \in \mathbb{R}, t \in [0,1]\}$ is such that $h(x,t)$ is integrable with respect to x for every t, and the family $\{\partial_t h(x,t) : x \in \mathbb{R}, t \in [0,1]\}$ of partial derivatives is uniformly integrable and continuous in t for every x, then $H(t) = \int h(x,t)\,dx$ is differentiable with

$$\frac{dH}{dt}(0) = \int \partial_t h(x,0)\,dx.$$

Here we consider eq. (9) and identify $h(\cdot,t) = K_t\, p_t$, so that

$$
\begin{aligned}
(11)\qquad \partial_t h(\cdot,t) &= (K_t + \partial_0 K_t)(p-q) \\
&\qquad + \left[(\partial_1 K_t)\left(\frac{p-q}{p_t}\right)' + (\partial_2 K_t)\left(\frac{p-q}{p_t}\right)''\right] p_t.
\end{aligned}
$$

Now $\partial_t h(\cdot,t)$ is continuous in t, because K and its partial derivatives are continuous, and their arguments depend continuously on t. Concerning uniform integrability, each of the terms K_t, $\partial_0 K_t$, $\partial_1 K_t$, $\partial_2 K_t$ grows at most polynomially as $x \to \pm\infty$. This is because by Lemma 4.2 and property (P4) of the class $\mathcal{P}$ the arguments of the terms grow at most polynomially; as $K \in \mathcal{R}_2$, the same is true for the functions themselves. Furthermore, by property (P3) and the above, the three factors in (11), namely $p - q$,

$$\left(\frac{p-q}{p_t}\right)' p_t = \left(\frac{p'-q'}{p_t} - \frac{p-q}{p_t}\frac{p_t'}{p_t}\right) p_t = p' - q' - (p-q)(\log p_t)',$$

and

$$\left(\frac{p-q}{p_t}\right)'' p_t \;=\; p''-q''-2\,(p'-q')\,(\log p_t)'-(p-q)\,(\log p_t)'' \\ +\,(p-q)\,((\log p_t)')^2,$$

decay faster than any polynomial as $x \to \pm\infty$. Therefore, the corresponding products in (11) with the factors involving K decay faster than any polynomial as well. By Lemma 4.2, this property holds uniformly in $t \in [0,1]$. Thus, the family (11) is uniformly integrable, and we may interchange the order of the integration and differentation. Similar growth and decay considerations show that the boundary terms in the partial integrations in (10) vanish. □

For use later on, we also state a version of Lemma 4.4, in which $K \in \mathcal{R}_1$ so that $\partial_2 K$ vanishes. The proof is analogous.

LEMMA 4.5. *Suppose that the kernel K depends on arguments x, z_0 and z_1 only and belongs to $\mathcal{R}_1$. Then the gradient G of the associated functional $\Phi_K : \mathcal{P} \to \mathbb{R}$ exists at any $q \in \mathcal{P}$, and is given by*

$$G = K + \partial_0 K - \frac{1}{q}\frac{d}{dx}[q\,\partial_1 K] \qquad (G = G_q,\ K = K_q). \tag{12}$$

By eqns. (8) and (12), a common form of the gradient valid for both $k = 1$ and $k = 2$ is $G = K + \partial_0 K + \Lambda K$, where ΛK is formally defined as the infinite sum

$$\Lambda K = \sum_{j=1}^{\infty} (-1)^j \, \frac{1}{q}\, \frac{d^j}{dx^j}[q\,\partial_j K],$$

where Λ and K depend tacitly on q. Of course, if $K \in \mathcal{R}_k$ all but the first k terms in the sum vanish, and so the definition makes good sense.

For the second step of the tangent construction let again $\Phi = \Phi_K$ be a kernel type functional associated with some kernel K in $\mathcal{R}_1$ or $\mathcal{R}_2$. If Φ is convex then the gradient G of Φ at $q \in \mathcal{P}$ is easily seen to be also a subgradient, and by Theorem 4.1 a proper score is obtained by setting

$$\begin{aligned} \mathrm{s} \;&=\; G - \int G\,q + \Phi(q) \\ &=\; K + \Lambda K + \partial_0 K - \int (\partial_0 K)\,q \qquad (\mathrm{s} = \mathrm{s}_q,\ \ K = K_q). \end{aligned} \tag{13}$$

As for the step leading to (13), note that by eqns. (8) and (12) we have

$$\Phi(q) - \int G\,q \;=\; \int K\,q - \left[\int K\,q + \int (\Lambda K)\,q \,+ \int (\partial_0 K)\,q\right]. \tag{14}$$

Then $\int (\Lambda K)\, q = 0$ because the integrand is a total derivative, the primitive of which vanishes as $x \to \pm\infty$ due to the growth and decay properties of the kernels in $\mathcal{R}_k$ and densities in $\mathcal{P}$. We will refer to this trivial observation as the *vanishing argument*. Since it will be used several times we state it below as a lemma, despite its simplicity. Finally, the first two integrals on the right-hand side of (14) cancel, leaving $-\int (\partial_0 K)\, q$ as the only remaining term and proving (13). This concludes the tangent construction of a proper score s from a convex functional $\Phi = \Phi_K$ where $K \in \mathcal{R}_k$ $(k = 1, 2)$.

LEMMA 4.6 (vanishing argument). *Let $q \in \mathcal{P}$, and let W be a real, differentiable function such that the function $g = q^{-1} \frac{d}{dx}\big[qW\big]$ is q-integrable, $\int |g| q < \infty$, and $\lim_{x\to\pm\infty} q(x)W(x) = 0$. Then $\int gq = 0$.*

Let us summarize the foregoing discussion.

PROPOSITION 4.7 (tangent construction). *Suppose that $K \in \mathcal{R}_k$ where $k = 1$ or $k = 2$, and that the associated functional Φ_K is convex. Then*

$$\mathrm{s} \;=\; K \;+\; \Lambda K \;+\; \partial_0 K \;-\; \int (\partial_0 K)\, q \qquad (\mathrm{s} = \mathrm{s}_q,\;\; K = K_q) \tag{15}$$

is a proper score relative to $\mathcal{P}$. It is local of order $2k$ if $\partial_0 K$ is constant in x for every $q \in \mathcal{P}$, or if $\int (\partial_0 K)\, q$ does not depend on q.

PROOF. The first claim has already been proved. Locality under the stated conditions is obvious: if $\partial_0 K = \partial_0 K(x, \log q(x), (\log q)'(x))$ (for $k = 1$, say) does not depend on x, it equals its expectation, $\int (\partial_0 K)\, q$. Finally, an explicit evaluation of the total differential(s) in the term ΛK yields partial derivatives of order $\le 2k$ only, thereby proving the order $2k$ claim. □

4.2. *Variational calculus.* A score s is proper if the functional $\mathcal{P} \ni p \mapsto \mathrm{S}(p, q)$ achieves its maximum at $p = q$, *for every* $q \in \mathcal{P}$. This circumstance allows a variational characterization of — in fact, a necessary condition for — the local proper scores.

LEMMA 4.8. *Suppose that $\mathrm{s} \in \mathcal{R}_2$ is a local proper score relative to $\mathcal{P}$. Then for every $q \in \mathcal{P}$ one has*

$$\partial_0 \mathrm{s} - \frac{1}{q}\frac{d}{dx}[q\, \partial_1 \mathrm{s}] + \frac{1}{q}\frac{d^2}{dx^2}[q\, \partial_2 \mathrm{s}] \;=\; c_q \qquad (s = s_q) \tag{16}$$

on $\mathbb{R}$, where

$$c_q \;=\; \int (\partial_0 \mathrm{s})\, q \qquad (s = s_q). \tag{17}$$

Proof. Fix $q \in \mathcal{P}$ and consider again convex combinations of the form $p_t = (1-t)q + tp$ where $p \in \mathcal{P}$ and $t \in (0,1)$. As the score is proper, we have $(\mathrm{S}(p_t,q) - \mathrm{S}(q,q))/t \le 0$ for every t. Let us compute the limit as $t \to 0$. Putting $\mathrm{s}_t = \mathrm{s}_{p_t}$ as before, we have at first

$$t^{-1}\,(\mathrm{S}(p_t,q) - \mathrm{S}(p_t,p_t)) \;=\; t^{-1}\int \mathrm{s}_t\,(q-p_t) \;=\; -\int \mathrm{s}_t\,(p-q).$$

Arguing in the same way as in the proof of Lemma 4.4, we find that the integrand is uniformly integrable and continuous in t, so the limit exists and equals $-\int \mathrm{s}\,(p-q)$ where $\mathrm{s} = \mathrm{s}_0 = \mathrm{s}_q$. Thus writing

$$\mathrm{S}(p_t,q) - \mathrm{S}(q,q) = \mathrm{S}(p_t,q) - \mathrm{S}(p_t,p_t) + \mathrm{S}(p_t,p_t) - \mathrm{S}(q,q)$$

and using Lemma 4.4 we get

$$\begin{aligned} &\lim_{t\to 0} t^{-1}\,(\mathrm{S}(p_t,q) - \mathrm{S}(q,q)) \\ &= \; -\int \mathrm{s}\,(p-q) \;+\; \int \left[\mathrm{s} + \partial_0\mathrm{s} - \frac{1}{q}\,\frac{d}{dx}[q\,\partial_1\mathrm{s}] + \frac{1}{q}\,\frac{d^2}{dx^2}[q\,\partial_2\mathrm{s}]\right](p-q) \\ &= \; \int (\partial_0\mathrm{s} + \Lambda\mathrm{s})\,(p-q). \end{aligned}$$

It follows that

$$\int (\partial_0\mathrm{s} + \Lambda\mathrm{s})\,(p-q) \;\le\; 0 \tag{18}$$

for $p \in \mathcal{P}$, which by the broadness of the class $\mathcal{P}$ is possible only if $\partial_0\mathrm{s} + \Lambda\mathrm{s}$ equals some constant c_q almost everywhere, hence everywhere by continuity. To prove constancy, let $f = \partial_0\mathrm{s} + \Lambda\mathrm{s}$ and $g = f - \int fq$. Then $\int gp \le 0$ for every $p \in \mathcal{P}$. Suppose g were not constant. Since $\int gq = 0$, the Lebesgue measure of the (open) set $\{g > 0\}$ is strictly positive. Thus, there exists a probability density $p_1 \in C^\infty$ with compact support such that $\int gp_1 > 0$. Then $p = (p_1 + q)/2 \in \mathcal{P}$ and $\int gp = \frac{1}{2}\int gp_1 > 0$, in contradiction to (18). Finally, the constant c_q is easily identified by integrating (16) against q and noting that $\int (\Lambda\mathrm{s})q = 0$, by the vanishing argument. □

Equation (16) essentially is the Euler equation of the calculus of variations (Gelfand and Fomin 1963, pp. 40–42). Its slightly different form here results from the fact that in our case the integrand of the functional to be optimized is of the form $F(x, \log y, (\log y)', (\log y)'')$ rather than of the common form $F(x, y, y', y'')$.

As a first application of the Euler equation we show that local proper scores are fixed points of the tangent construction. To this end, let $\mathrm{s} \in \mathcal{R}_2$ be a local proper score of order two. By Theorem 4.1 and Lemma 4.4, the functional $\Phi_{\mathrm{s}} : \mathcal{P} \to \mathbb{R}$ associated with the kernel s is convex. The tangent construction then gives

$$\widetilde{\mathrm{s}} \;=\; \mathrm{s} + \Lambda \mathrm{s} + \partial_0 \mathrm{s} - \int (\partial_0 \mathrm{s})\, q \tag{19}$$

on substituting s for K in Proposition 4.7. Initially, this is another proper score, possibly of higher order, and possibly nonlocal. However, by Lemma 4.8 the right-hand side of (19) reduces to s, whence in fact $\widetilde{\mathrm{s}} = \mathrm{s}$.

Proposition 4.9. *For a local proper score* $\mathrm{s} \in \mathcal{R}_2$ *the tangent construction based on the (convex) functional* Φ_{s} *leads back to* s*. That is, any local proper score of order two is a fixed point of the tangent construction.*

4.3. *Construction of a z_2-independent kernel.*

The vanishing argument of Lemma 4.6 enables us to modify a given kernel without changing the associated functional. This strategy is utilized in the following explicit construction of a z_2-independent kernel from a given local proper score. Again, it is tacitly assumed that the, as arguments often suppressed, quantities z_j refer to a fixed density $q \in \mathcal{P}$, that is, $z_j = z_j(q, x)$.

Proposition 4.10. *Given the local proper score* $\mathrm{s} \in \mathcal{R}_2$*, let the kernel* K *be defined as*

$$K \;=\; \mathrm{s} - \frac{1}{q}\,\frac{d}{dx}[q\,V] \;=\; \mathrm{s} - \left[z_1 + \frac{d}{dx} \right] V \tag{20}$$

where

$$V \;=\; \int_0^{z_1} \partial_2 \mathrm{s}(x, z_0, t, z_2)\, dt. \tag{21}$$

Then $K \in \mathcal{R}_1$ *and* $\Phi_K = \Phi_{\mathrm{s}}$*. In particular, the score* s *can be reconstructed from the kernel* K *via the tangent construction.*

Proof. The kernel K clearly inherits the polynomial growth properties from s, and it is twice continuously differentiable since $\mathrm{s} \in C^4$. In particular, K is well-defined. An application of the vanishing argument to the term $\frac{1}{q}\frac{d}{dx}[q\,V]$ shows that the kernels K and s give rise to the same functional. Thus $\Phi_K = \Phi_{\mathrm{s}}$, and the last claim follows from Propositions 4.7 and 4.9.

Therefore, to complete the proof it remains to show that the kernel K from (20) does not depend on z_2, or else, that $\partial_2 K = 0$.

A comparison of the two differential operators

$$\begin{aligned}\partial_2 \frac{d}{dx} &= \partial_2 \left(\partial_x + z_1\partial_0 + z_2\partial_1 + z_3\partial_2\right) \\ &= \partial^2_{x2} + z_1\partial^2_{02} + \partial_1 + z_2\partial^2_{12} + z_3\partial^2_{22}\,, \\ \frac{d}{dx}\partial_2 &= \partial^2_{x2} + z_1\partial^2_{02} + z_2\partial^2_{12} + z_3\partial^2_{22}\end{aligned}$$

yields the commutation relation

$$\partial_2 \frac{d}{dx} = \frac{d}{dx}\partial_2 + \partial_1. \tag{22}$$

Therefore, using $\partial_1 V = \partial_2 \mathrm{s}$ (see (21)) we get

$$\partial_2 K = \partial_2 \mathrm{s} - z_1 \partial_2 V - \partial_1 V - \frac{d}{dx}\partial_2 V = -\left[z_1 + \frac{d}{dx}\right]\partial_2 V,$$

where

$$\partial_2 V = \int_0^{z_1} \partial^2_{22}\mathrm{s}(x, z_0, t, z_2)\, dt. \tag{23}$$

Thus if

$$\partial^2_{22}\mathrm{s}(x, z_0(q,x), t, z_2(q,x)) = 0 \quad \text{for all } x \in \mathbb{R},\ q \in \mathcal{P},\ |t| \le |z_1(q,x)|, \tag{24}$$

then $\partial_2 V = 0$ and hence $\partial_2 K = 0$, that is $K \in \mathcal{R}_1$, as claimed. The somewhat lengthy proof of (24) is given in the next subsection.

4.4. *Proof of the independence condition (24).* The proof primarily rests upon the Euler equation (16). An evaluation of the total derivatives in (16) shows at first that the Euler equation can be written in the form

$$z_0^{(4)} \cdot a(x, z_0, z_0', z_0'', z_0''') - b(x, z_0, z_0', z_0'', z_0''') = c_q \quad (z_0 = \log q) \tag{25}$$

with

$$a = \partial^2_{22}\mathrm{s} \qquad (\text{so that in fact } a = a(x, z_0, z_0', z_0''))$$

and a function b, which depends (only) on the scoring function s and its partial derivatives up to order three, other than x, z_0, and the logarithmic derivatives z_1, z_2 and $z_3 = z_2'$. Therefore, and because s is of smoothness class C^4, the function b is continuously differentiable. The same holds for $a = \partial^2_{22}\mathrm{s}$, of course, so that a, $b \in C^1$.

We will at first show that the constant c_q is in fact independent of q. The ensuing fact that one and the same equation, (25) with $c_q = c$, holds for every $q \in \mathcal{P}$ is then utilized to complete the proof of (24). For each of these steps it is important that the class $\mathcal{P}$ is sufficiently rich. Let

$$Z(p, x, k) = (z_0(p, x), z_1(p, x), \ldots, z_k(p, x))$$

for $k = 0, 1, \ldots$ with $z_j(p, x) = (\log p)^{(j)}(x)$ as above.

LEMMA 4.11 ($\mathcal{P}$-richness). *Let $k \in \{0, 1, \ldots, 4\}$. (a) For every $x \in \mathbb{R}$ and $y \in \mathbb{R}^{k+1}$ there exists $p \in \mathcal{P}$ such that $Z(p, x, k) = y$. (b) For every pair p_1, $p_2 \in \mathcal{P}$ there exist $p \in \mathcal{P}$ and $x \in \mathbb{R}$ such that $Z(p, x, k) = Z(p_1, x, k)$ and $p(u) = p_2(u)$ for u outside some neighborhood of x.*

PROOF. This is fairly obvious from the definition of $\mathcal{P}$. For completeness, we include a proof. As concerns part (a), let $x \in \mathbb{R}$ and $y \in \mathbb{R}^{k+1}$ be fixed. There is some $q \in \mathcal{P}$ such that $z_0(q, x) = y_0$. We will construct p as a perturbation $p = q(1 + \psi)$ of q such that $\psi(x) = 0$. Certainly $p \in \mathcal{P}$ if ψ has compact support and is such that $1 + \psi > 0$, $\int \psi q = 0$, and $\psi \in C^4$. It suffices to show that it is possible to prescribe arbitrary values for the first k derivatives of ψ at x, subject to those conditions. To that end, let $\psi = PM$ where

$$P(u) = \sum\nolimits_{j=1}^{r} a_j (u - x)^j$$

is a polynomial vanishing at x and $0 \le M \in C^4$ is a mollifier type function with (small) compact support S such that $M(x) = 1$ and $M^{(j)}(x) = 0$ for $j = 1, \ldots, k$. Then $\psi^{(j)}(x) = P^{(j)}(x)$ $(j = 1, \ldots, k)$, confirming that arbitrary values can indeed be prescribed if $r \ge k$. By increasing r if necessary, one can further assume that P attains both positive and negative values on S.

Let any such P be fixed. We show that the conditions $PM > -1$ and $\int PMq = 0$ can be satisfied, too. In fact, since $P(x) = 0$ one can modify M such that $PM > -1$ everywhere without affecting its local behavior at x. Since $\int_{\{P<0\}} PMq < 0$ there is $\delta > 0$ such that the interval $J = [x - \delta,\ x + \delta]$ is contained in the interior of S and $\int_J PMq + \int_{J^c \cap \{P<0\}} PMq < 0$. Finally, on the set $S \cap J^c \cap \{P > 0\}$ one can modify M such that

$$\int PMq \;=\; \int_J PMq \;+ \int_{J^c \cap \{P<0\}} PMq \;+ \int_{J^c \cap \{P>0\}} PMq \;=\; 0,$$

without affecting the condition $PM > -1$. This concludes the proof of (a). For part (b), note that because p_1, p_2 are continuous probability densities, there is $x \in \mathbb{R}$ such that $p_1(x) = p_2(x)$. The above construction then yields a local perturbation $p \in \mathcal{P}$ of p_2 satisfying $Z(p, x, k) = Z(p_1, x, k)$. □

LEMMA 4.12. *The constant c_q in (25) does not depend on the density q: there is a constant $c \in \mathbb{R}$ such that $c_q = c$ for every $q \in \mathcal{P}$.*

PROOF. We use an argument in Section 4 of Parry et al. (2011). Eq. (25) (resp. (16)) can be condensed to a statement of the form

$$F(x, Z(q, x, 4)) = c_q \qquad (x \in \mathbb{R},\ q \in \mathcal{P}), \tag{26}$$

where the function F is determined by the score s alone. Suppose that c_q is not independent of q. Then there are $q_1,\, q_2 \in \mathcal{P}$ such that $c_{q_1} \neq c_{q_2}$. By Lemma 4.11 (b) there exist $q \in \mathcal{P}$ and $x_1 \neq x_2 \in \mathbb{R}$ such that $Z(q, x_1, 4) = Z(q_1, x_1, 4)$ and $Z(q, x_2, 4) = Z(q_2, x_2, 4)$. By (26) it follows that for both $j = 1$ and $j = 2$

$$c_{q_j} = F(x_j, Z(q_j, x_j, 4)) = F(x_j, Z(q, x_j, 4)) = c_q.$$

The contradiction implies that c_q is indeed independent of q. □

The following lemma is an easy consequence of the uniqueness theorem for higher-order differential equations.

LEMMA 4.13 (reduction principle). *Let $k \in \{0, 1, 2, 3\}$, and let a and b be C^1 functions of arguments $x, y_0, \ldots, y_k$. Suppose that the function $z_0 = \log q(x)$, $x \in \mathbb{R}$ is, for every $q \in \mathcal{P}$, a solution of the differential equation*

$$z^{(k+1)} \cdot a(x, z, \ldots, z^{(k)}) = b(x, z, \ldots, z^{(k)}). \tag{27}$$

Then $a(x, Z(q, x, k)) = 0$ for every $x \in \mathbb{R}$, $q \in \mathcal{P}$.

PROOF. Fix $q \in \mathcal{P}$ and $x \in \mathbb{R}$, and suppose that $a(x, Z(q, x, k)) \neq 0$. Then there is an open interval containing x on which $a(\cdot, Z(q, \cdot, k))$ does not vanish and $b(\cdot, Z(q, \cdot, k))/a(\cdot, Z(q, \cdot, k))$ is continuosly differentiable. Therefore $\log q$ is, perhaps in a smaller neighborhood of x, the only solution to the equation (27) whose derivatives up to order k at x are given by the components of the vector $Z(q, x, k)$. On the other hand, by Lemma 4.11 (a) there exists $p \in \mathcal{P}$ such that $Z(p, x, k) = Z(q, x, k)$ but $z_{k+1}(p, x) \neq z_{k+1}(q, x)$. By assumption this p, too, is a solution of (27), with the same initial conditions. This contradiction to uniqueness is resolved only if $a(x, Z(q, x, k)) = 0$. Since $q \in \mathcal{P}$ and $x \in \mathbb{R}$ were arbitrary, the proof of the lemma is complete. □

Let us combine these facts. Absorbing the (universal) constant $c_q = c$ in (25) into the function b we see that every $q \in \mathcal{P}$ satisfies a differential equation of the form (27) with $a = \partial^2_{22}\mathrm{s}$. Therefore $\partial^2_{22}\mathrm{s}(x, Z(q, x, 2)) = 0$ for all $q \in \mathcal{P}$ and $x \in \mathbb{R}$ by the reduction principle. The proof of (24) is completed on noting that for any $x \in \mathbb{R}$, $q \in \mathcal{P}$, $|t| \leq |z_1(q, x)|$ there is $p \in \mathcal{P}$ such that $Z(p, x, 2) = (z_0(q, x), t, z_2(q, x))$, by Lemma 4.11 (a).

4.5. *Linear dependence on z_0.* The fact that a local proper score $\mathrm{s} \in \mathcal{R}_2$ can be represented by means of a z_2-independent kernel $K \in \mathcal{R}_1$ will now be utilized to show that both s and K depend linearly on the logarithmic score, z_0.

PROPOSITION 4.14 (linearity in z_0). *Let $K \in \mathcal{R}_1$ be the kernel constructed in Section 4.3 from a given local proper score $\mathrm{s} \in \mathcal{R}_2$. Then K is of the form $K(x, z_0, z_1) = cz_0 + K_0(x, z_1)$ where c is a real constant, and s is of the form (4) (with the same c).*

PROOF. We already know that the score s can be represented as in (15), with K not depending on z_2. Furthermore, by the Euler equation (16) and Lemma 4.12 there is some constant c such that

$$\partial_0 \mathrm{s} - c = \frac{1}{q}\frac{d}{dx}\left(q\partial_1 \mathrm{s} - \frac{d}{dx}\big[q\partial_2 \mathrm{s}\big]\right). \tag{28}$$

Using these facts along with the commutation relation $\partial_1 \frac{d}{dx} = \frac{d}{dx}\partial_1 + \partial_0$ (cf. (22)) we get

$$\begin{aligned}
\partial_1 \mathrm{s} &= \partial_1 K - \partial_1 K - z_1 \partial_{11}^2 K - \partial_1\left(\frac{d}{dx}\big[\partial_1 K\big]\right) + \partial_{01}^2 K \\
&= -z_1 \partial_{11}^2 K - \frac{d}{dx}\big[\partial_{11}^2 K\big] - \partial_{01}^2 K + \partial_{01}^2 K \\
&= -z_1 \partial_{11}^2 K - \frac{d}{dx}\big[\partial_{11}^2 K\big].
\end{aligned}$$

On the other hand, we have

$$\partial_2 \mathrm{s} = -\partial_2 \frac{d}{dx}\partial_1 K = -\partial_2\left(\partial_{x1}^2 K + \partial_{01}^2 K \cdot z_1 + \partial_{11}^2 K \cdot z_2\right) = -\partial_{11}^2 K,$$

and hence

$$q^{-1}\frac{d}{dx}\big[q\,\partial_2 \mathrm{s}\big] = -z_1\,\partial_{11}^2 K - \frac{d}{dx}\big[\partial_{11}^2 K\big].$$

Thus $q\partial_1 \mathrm{s} = \frac{d}{dx}\big[q\,\partial_2 \mathrm{s}\big]$, the right-hand side of (28) vanishes, and $\partial_0 \mathrm{s}$ is constant, $\partial_0 \mathrm{s} = c$. It follows that $\mathrm{s} = cz_0 + g(x, z_1, z_2)$ for some function g independent of z_0, and it remains to verify the particular forms of K and s.

By (20) and the special form of s we have $K - cz_0 = g - z_1 V - \frac{d}{dx}V$ where now $V = \int_0^{z_1} \partial_2 g(x, t, z_2)\, dt$. But $\partial_2 V = 0$, by (23) and (24), and clearly $\partial_2(K - cz_0) = 0$, since $K \in \mathcal{R}_1$. Therefore $K_0 = g - z_1 V - \frac{d}{dx}V$ does not depend on z_2, and it also does not depend on z_0 (since neither g nor V depend on z_0), so $K_0 = K_0(x, z_1)$. This completes the proof of the first claim. The tangent construction based on $K = cz_0 + K_0(x, z_1)$ then implies, upon observing $\partial_0 K - \int (\partial_0 K) q = 0$, that

$$\begin{aligned} \mathrm{s} &= K - z_1\, \partial_1 K - \frac{d}{dx}\left[\partial_1 K\right] \\ &= cz_0 + K_0 - z_1\, \partial_1 K_0 - \partial^2_{x1} K_0 - z_2\, \partial^2_{11} K_0, \end{aligned}$$

which is the desired representation. □

4.6. *Completion of the proof of Theorem 3.2.* The tangent construction based on a convex functional Φ_K with $K \in \mathcal{R}_1$ yields a proper score, which is of the form (4) if the kernel K is of the form (3). This proves part (a). Part (b) follows from Propositions 4.10 and 4.14. Finally, part (c) is immediate from Theorem 4.1.

5. Remaining proofs, supplements, and examples.

5.1. *Proof of Proposition 3.3.* Initially, suppose that $K \in \mathcal{R}_1$ does not depend on z_0, so that $K = K(x, z_1)$, and is convex in z_1 for every fixed x. Given p_0, $p_1 \in \mathcal{P}$ and $t \in [0, 1]$, let $p_t = tp_1 + (1 - t)p_0$ and put $\alpha = tp_1/p_t$, pointwise for every $x \in \mathbb{R}$. Then $p_t'/p_t = \alpha\, p_1'/p_1 + (1 - \alpha)\, p_0'/p_0$, whence

$$K\left(\cdot\,, p_t'/p_t\right) \le \alpha K\left(\cdot\,, p_1'/p_1\right) + (1 - \alpha) K\left(\cdot\,, p_0'/p_0\right),$$

and so

$$\begin{aligned} \Phi_K(p_t) &= \int K\left(x, \frac{p_t'}{p_t}(x)\right) p_t(x)\, dx \\ &\le \int \left[\alpha(x) K\left(x, \frac{p_1'}{p_1}(x)\right) + (1 - \alpha(x)) K\left(x, \frac{p_0'}{p_0}(x)\right)\right] p_t(x)\, dx \\ &= \int K\left(x, \frac{p_1'}{p_1}(x)\right) tp_1(x)\, dx + \int K\left(x, \frac{p_0'}{p_0}(x)\right) (1 - t) p_0(x)\, dx \\ &= t\Phi_K(p_1) + (1 - t)\Phi_K(p_0). \end{aligned}$$

The general case follows by the strict convexity of the entropy functional $p \mapsto \int p \log p$. Concerning the claim about strict propriety, the pathology described in Remark 3.8 does not occur within the class $\mathcal{P}$, because all densities $p \in \mathcal{P}$ are strictly positive. Thus, the primitive of p'/p exists throughout $\mathbb{R}$ and equals $\log p$ up to a constant, so that $p'/p = q'/q$ implies $p = q$. □

5.2. *Local proper scoring rules of order one.* The representation (4) suggests that local proper scores of exact order $k = 1$ do not exist. In fact, Parry et al. (2011) have shown that there are no key local score functions of odd order. Within our framework, we can prove the following.

PROPOSITION 5.1. *Any local score* $\mathrm{s} \in \mathcal{R}_1$ *that is proper relative to* $\mathcal{P}$ *is of the form* $\mathrm{s} = cz_0 + k(x)$ *for some* $c \geq 0$.

PROOF. Suppose that $\mathrm{s} \in \mathcal{R}_1$ is proper. The Euler equation reduces to

$$\partial_0 \mathrm{s} - \frac{1}{q}\frac{d}{dx}[q\,\partial_1 \mathrm{s}] \;=\; \partial_0 \mathrm{s} + z_1\partial_1 \mathrm{s} - \partial^2_{x1}\mathrm{s} - z_1\partial^2_{01}\mathrm{s} - z_2\partial^2_{11}\mathrm{s} \;=\; c_q \;\;(s = s_q)$$

in this case. Arguing as in Section 4.4, we find that $c_q = c$ is independent of q and that $\partial^2_{11}\mathrm{s}$ vanishes on $\mathbb{R}^3$. Therefore there are functions g, h depending only on x, z_0 such that $\mathrm{s} = z_1 g + h$. Plugging this representation into the Euler equation gives

$$c \;=\; z_1\partial_0 g + \partial_0 h + z_1 g - \partial_x g - z_1\partial_0 g \;=\; z_0' g - \partial_x g + \partial_0 h,$$

whence $g = 0$ by another application of the reduction principle. Thus $\partial_0 h = c$, which means that $\mathrm{s} = cz_0 + k(x)$. Since z_0 represents the logarithmic score, s can be proper only if $c \geq 0$. □

5.3. *Examples.* In the subsequent examples, we keep the notation to a minimum and suppress arguments whenever possible.

EXAMPLE 5.2. For $n \geq 2$ even and $c \geq 0$, let $K = cz_0 + z_1^n$. Then $K \in \mathcal{R}_1$, the functional Φ_K is stricly convex on $\mathcal{P}$, and the tangent construction of Proposition 4.7 yields the score

$$\begin{aligned}
\mathrm{s} \;&=\; K - z_1\partial_1 K - \frac{d}{dx}\partial_1 K + \partial_0 K - \int(\partial_0 K)q \\
&=\; cz_0 + z_1^n - nz_1^n - n(n-1)z_1^{n-2}z_2 + c - c \\
&=\; cz_0 - (n-1)\left(z_1^n + nz_1^{n-2}z_2\right),
\end{aligned}$$

which is local of order two and strictly proper relative to $\mathcal{P}$.

Conversely, if s is as above, let us carry out the construction of the associated kernel K described in Section 4.3. We set

$$V \;=\; \int_0^{z_1} \partial_2 \mathrm{s}(x, z_0, t, z_2)\,dt = -nz_1^{n-1}$$

and then define K as

$$\begin{aligned} K &= \mathrm{s} - \left[z_1 + \frac{d}{dx}\right] V \\ &= \mathrm{s} + n z_1^n + n(n-1) z_1^{n-2} z_2 \\ &= c z_0 - (n-1)\left(z_1^n + n z_1^{n-2} z_2\right) + n z_1^n + n(n-1) z_1^{n-2} z_2 \\ &= c z_0 + z_1^n. \end{aligned}$$

The construction indeed recovers the kernel K from the score s.

Example 5.3. The special case $K = z_1^2$ in the previous example gives the Hyvärinen score, $\mathrm{s} = -z_1^2 - 2z_2$. Being quadratic in the log likelihood derivative, $z_1 = q'/q$, and linear in the second derivative, $z_2 = q''/q - (q'/q)^2$, this score generally is sensitive to outliers. For example, within the Gaussian shift-scale family with mean μ and variance σ^2, the Hyvärinen score reduces to $\mathrm{s} = -(x-\mu)^2/\sigma^4 + 2/\sigma^2$.

As an alternative, let us consider the kernel $K = \log\cosh z_1$, which grows only linearly as z_1 becomes large. The corresponding score

$$\mathrm{s} = \log\cosh z_1 - z_1 \tanh z_1 - z_2\left(1 - \tanh^2 z_1\right) \tag{29}$$

appears to be more robust, because as $|y| \to \infty$,

$$\log\cosh y - y\,\tanh y \to \log(1/2),$$

and the factor of z_2 tends to zero exponentially, in that $1 - \tanh^2 y \sim 4\exp\left(-2|y|\right)$. Of course, the log cosh score (29) is strictly proper relative to $\mathcal{P}$, since K is strictly convex.

6. Data example: Probabilistic weather forecasting.

The data example in this section illustrates the use of local and non-local scoring rules in an applied forecasting problem.

Weather forecasting has traditionally been viewed as a deterministic enterprise that draws on highly sophisticated, numerical models of the atmosphere. The advent of ensemble prediction systems in the early 1990s marks a change of paradigms towards probabilistic forecasting (Palmer 2002; Gneiting and Raftery 2005). An ensemble prediction system consists of multiple runs of numerical weather prediction models, which differ in the initial conditions and/or the mathematical representation of the atmosphere. As ensemble forecasts are subject to dispersion errors and biases, some form of statistical postprocessing is required, for a happy marriage of mechanistic and statistical modeling.

Wilks and Hamill (2007) and Bröcker and Smith (2008) review statistical postprocessing techniques for ensemble weather forecasts. State of the art methods include the Bayesian model averaging (BMA) approach developed by Raftery et al. (2005) and Sloughter et al. (2007, 2010), and the heterogeneous regression, or ensemble model output statistics (EMOS), technique of Gneiting et al. (2005) and Thorarinsdottir and Gneiting (2010). The BMA approach employs a mixture distribution, where each mixture component is a parametric probability density associated with an individual ensemble member, with the mixture weight reflecting the member's relative contributions to predictive skill over a training period. In contrast, the EMOS predictive distribution is a single parametric distribution.

For concreteness, consider an ensemble of point forecasts, $f_1, \ldots, f_k$, for surface temperature, x, at a given time and location. The goal is to fit predictive distributions that are as sharp as possible, subject to them being calibrated (Gneiting, Balabdaoui and Raftery 2007). Let $\phi(x; \mu, \sigma^2)$ denote the normal density with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ evaluated at $x \in \mathbb{R}$. The BMA approach of Raftery et al. (2005) employs Gaussian components with a linearly bias-corrected mean. The BMA predictive density for temperature then becomes

$$p(x \,|\, f_1, \ldots, f_k) \; = \; \sum_{i=1}^{k} w_i \, \phi(x; a_i + b_i \, f_i, \, \sigma^2),$$

with BMA weights, $w_1, \ldots, w_k$, that are nonnegative and sum to 1, bias parameters $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$, and a common variance parameter, σ^2. The EMOS approach of Gneiting et al. (2005) employs a single Gaussian predictive density, in that

$$p(x \,|\, f_1, \ldots, f_k) = \phi(x; a + b_1 \, f_1 + \cdots + b_k \, f_k, \, c + d \, s^2),$$

with regression parameters a and $b_1, \ldots, b_k$, and spread parameters c and d, where s^2 is the variance of the ensemble values. The EMOS technique thus is more parsimonious, while the BMA method is more flexible.

Following the original development in Raftery et al. (2005) and Gneiting et al. (2005), we apply the BMA and EMOS methods to the five-member University of Washington Mesoscale Ensemble over the North American Pacific Northwest (Grimit and Mass 2002), at a prediction horizon of 48 hours. Here we compare the predictive performance of the BMA and EMOS density forecasts for surface temperature verifying in the period of 24 April to 30 June 2000, which is the largest period common to those used by Raftery et al. (2005) and Gneiting et al. (2005). The predictive models were fitted on

Table 1

Mean logarithmic score (LS), Hyvärinen score (HS), log cosh score (LCS), quadratic score (QS) and spherical score (SphS) for statistically postprocessed ensemble forecasts of surface temperature over the North American Pacific Northwest in April–June 2000, using Bayesian model averaging (BMA) and ensemble model output statistics (EMOS), respectively. See the text for details.

Scoring Rule	LS	HS	LCS	QS	SphS
BMA	−2.502	0.113	0.0572	0.101	0.319
EMOS	−2.486	0.118	0.0595	0.103	0.321

trailing training periods of length 25 days for BMA and length 40 days for EMOS, as recommended and described in the aforementioned papers. Overall, there were 23,691 individual forecast cases at individual meteorological stations and valid times, when aggregated temporally and spatially over the test period and the Pacific Northwest, comprising the states of Washington, Oregon and Idaho, and the southern part of the Canadian province of British Columbia. All scores reported are averaged over the 23,691 forecast cases.

In Table 1 we assess these forecasts, by computing the mean score under various local proper scoring rules, namely the logarithmic score (LS), the Hyvärinen score (HS) and the log cosh score (LCS) introduced in eq. (29). In addition, we consider two popular non-local scores, namely the quadratic score (QS) and the spherical score (SphS), defined as

$$\mathrm{QS}(p,x) = 2\,p(x) - \|p\|_2^2 \qquad \text{and} \qquad \mathrm{SphS}(p,x) = \frac{p(x)}{\|p\|_2},$$

respectively, where $\|\cdot\|_2$ denotes the L_2-norm. These scores are strictly proper relative to the class of the probability measures with square-integrable Lebesgue densities (Matheson and Winkler 1976; Gneiting and Raftery 2007).

Under all scoring rules, the EMOS technique shows a slightly higher (that is, better) mean score than the BMA method. However, the differences pale when compared to those between the unprocessed ensemble forecast and the statistically postprocessed density forecasts. The unprocessed five-member ensemble gives a discrete predictive distribution, namely the empirical measure in $f_1, \ldots, f_5$, to which the above scores do not apply directly. However, we can compute the mean score for a smoothed ensemble forecast, which we take to be normal, with the first two moments identical to those of the empirical measure. Under this natural approach, the mean scores for the smoothed ensemble forecast are very low, reaching −21.4 for the logarithmic score, -1.14×10^4 for the Hyvärinen score, −0.230 for the log cosh score,

-0.194 for the quadratic score, and 0.217 for the spherical score, thereby attesting to the benefits of statistical postprocessing.

7. Discussion.

A scoring rule on the real line is local of order k if the score depends on the predictive density only through its value, and the values of its derivatives of order up to k, at the realizing event. In this paper we have characterized the class of the local proper scoring rules of order $k \leq 2$. A practically useful characterization depends on the judicious choice of a class $\mathcal{S}$ of scoring functions, and a class $\mathcal{D}$ of predictive densities, within which scores and densities may vary freely. Involved therein is a delicate tradeoff, in that narrow classes $\mathcal{D}$ allow for weak assumptions on the members of $\mathcal{S}$, but have little, if any, practical relevance. Our choice of $\mathcal{S}$ — the class $\mathcal{R}_2$ of scoring functions growing at most polynomially at infinity — and of $\mathcal{D}$ — the class $\mathcal{P}$ of densities decaying faster than any polynomial, with log likelihood derivatives growing at most polynomially — appears to be usefully general and achieving a reasonable balance. The balance could easily be shifted, for example, in favor of more heavy-tailed densities, by adapting the polynomial growth order in $\mathcal{S}$.

Counterexamples show that proper scoring rules of practical interest, such as the Hyvärinen score (2), may no longer be strictly proper relative to any class $\mathcal{D}$ that contains a convex family of densities with a single common zero. It is thus natural to assume that all densities in $\mathcal{D}$ are strictly positive on their common support, Ω, which then is an interval. The case of finite boundary points, for example when $\Omega = (0, \infty)$, appears to be tractable similarly to the case $\Omega = \mathbb{R}$ considered here, and resulting in essentially the same characterization. It suffices to impose suitable boundary conditions at $x = 0$ on the classes $\mathcal{S}$ and $\mathcal{D}$, guaranteeing the existence of integrals and causing the boundary terms in the proof of Lemma 4.4 to vanish.

In a recent *tour de force*, Parry, Dawid and Lauritzen (2011) investigated local proper scoring rules on the real line of any order $k \geq 0$. In an elegant approach based on operator algebra, they established the existence of key local score functions for any even order, and their non-existence for odd orders, in addition to studying their invariance under data transformations. In the case $k = 2$ the general form of the local proper scoring rules in Parry et al. (2011) is essentially equivalent to ours, up to the parametrization in terms of densities rather than log densities. However, there are important differences, including the basic approach and techniques employed.

A key local score derives from the homogeneous Euler-Lagrange equation, which characterizes the scores for which every density q is a stationary point of the mapping $p \mapsto \mathrm{S}(p, q)$. Accordingly, Parry et al. (2011)'s analysis is in terms of differential calculus, which leads to separate discussions of the boundary terms from partial integrations and of sufficient conditions for (strict) propriety. The latter occur in the form of convexity conditions on homogeneous q-functions, which correspond to our kernels; Proposition 3.3 states essentially the same result in the case $k = 2$.

In a different ansatz, our work starts from the characterization of proper scoring rules via convex functionals and their (sub-)gradients (Hendrickson and Buehler 1971; Gneiting and Raftery 2007). This readily yields the basic form (15) of the second-order local proper scoring rules in a natural tangent construction, up to a possibly nonlocal term. Only then we apply the calculus of variations to show that the possibly nonlocal term vanishes, which establishes the definite form (4). Control of the boundary terms from partial integrations is vital, and is achieved through our particular choice of the classes of scoring functions and predictive densities. The explicit specification of the classes $\mathcal{S}$ and $\mathcal{D}$, along with the tangent construction, allow us to give a rigorous, yet full-fledged and practically relevant characterization of the second-order local proper scoring rules, hence constitute the main original contributions of our work.

With the resurgence of interest in probabilistic forecasting (Gneiting 2008), scoring rules for density forecasts are in increasing demand. In this context, locality is an appealing property, which we have studied in this work. A different argument posits that a scoring rule for probabilistic forecasts ought to be sensitive to distance, in the sense that it rewards not just the assignment of greater mass to exactly the event or value that is observed, but also to nearby events (Staël von Holstein 1969; Jose, Nau and Winkler 2009). While either approach has appeal, locality and sensitivity to distance appear to be mutually exclusive properties, and it is not clear which one is more compelling (Mason 2008; Winkler and Jose 2008). However, in our meteorological data example as well as in other experience, local and non-local proper scoring rules generally yield comparable results.

In addition to their use in the assessment of predictive performance, proper scoring rules play major roles in the theory and practice of estimation (Dawid 2007; Gneiting and Raftery 2007). A striking aspect is that local proper scoring rules of order $k \geq 2$ allow for statistical inference without knowledge of normalization constants (Parry et al. 2011). Indeed, this was the motivation for the initial development by Hyvärinen (2005). The example of the log cosh score (29) shows that local scores can be less nonrobust than one might expect. These facets suggest exciting opportunities and novel prospects particularly in complex settings. Undoubtedly, the pioneering work of Hyvärinen (2005, 2007), Dawid and Lauritzen (2005) and

Parry et al. (2011) has laid the groundwork for a wide range of promising future work, both theoretically and methodologically, and including discrete and multivariate settings, where our tangent approach may continue to be useful and provide new insight.

REFERENCES

Bauer, H. (2001). *Measure and Integration Theory.* W. de Gruijter, Berlin.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, **7**, 686–690.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Bröcker, J. and Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus Ser. A*, **60**, 663–678.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer, New York.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Ser. A*, **147**, 278–292.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, **59**, 77–93.

Dawid, A. P (2008). Comments on: Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test*, **17**, 243–244.

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, University of Tokyo, pp. 22–28.

Dawid, A. P., Parry, M. and Lauritzen, S. (2009). Personal communication.

Ehm, W. and Gneiting, T. (2009, Addendum 2010). Local proper scoring rules. University of Washington, Department of Statistics, Technical Report no. 551.

Gelfand, I. M. and Fomin, S. V. (1963). *Calculus of Variations.* Prentice-Hall, Englewood Cliffs, New Jersey.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality.* Chapman & Hall/CRC, Boca Raton.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society Ser. A*, **171**, 319–321.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, **310**, 248–249.

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

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Ser. B*, **69**, 243–268.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005).

Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Ser. B*, **14**, 107–114.

Grimit, E. P. and Mass, C. F. (2002). Initial results of a mesoscale short-range ensemble system over the Pacific Northwest. *Weather and Forecasting*, **17**, 192–205.

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *Annals of Mathematical Statistics*, **42**, 1916–1921.

Huber, P. J. (1974). Fisher information and spline interpolation. *Annals of Statistics*, **2**, 1029–1033.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499–2512.

Jose, V. R. R., Nau, R. F. and Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science*, **55**, 582–590.

Mason, S. J. (2008). Understanding forecast verification statistics. *Meteorological Applications*, **15**, 31–40.

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

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, **128**, 747–774.

Parry, M., Dawid, A. P. and Lauritzen, S. (2011). Proper local scoring rules. Preprint, arXiv:1101.5011v1.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010). Probabilistic wind forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, **105**, 25–35.

Staël von Holstein, C.-A. S. (1969). A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology*, **9**, 360–364.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Ser. A*, **173**, 371–388.

Villani, C. (2009). *Optimal Transport*. Grundlehren der mathematischen Wissenschaften, Vol. 338. Springer, Berlin.

Wilks, D. S. and Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, **135**, 2379–2390.

Winkler, R. L. and Murphy, A. H. (1968). 'Good' probability assessors. *Journal of Applied Meteorology*, **7**, 751–758.

Winkler, R. L. and Jose, V. R. R (2008). Comments on: Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test*, **17**, 251–255.

Institute for Frontier Areas of Psychology
and Mental Health
Wilhelmstr. 3a
79098 Freiburg
Germany

Institute for Applied Mathematics
University of Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
Germany