The Annals of Statistics
2012, Vol. 40, No. 1, 561–592
DOI: 10.1214/12-AOS971

PROPER LOCAL SCORING RULES

BY MATTHEW PARRY[1], A. PHILIP DAWID AND STEFFEN LAURITZEN

University of Otago, University of Cambridge and University of Oxford

We investigate proper scoring rules for continuous distributions on the real line. It is known that the log score is the only such rule that depends on the quoted density only through its value at the outcome that materializes. Here we allow further dependence on a finite number m of derivatives of the density at the outcome, and describe a large class of such m-*local* proper scoring rules: these exist for all even m but no odd m. We further show that for $m \geq 2$ all such m-local rules can be computed without knowledge of the normalizing constant of the distribution.

1. Introduction. A *scoring rule* $S(x, Q)$ is a loss function measuring the quality of a quoted distribution Q, for an uncertain quantity X, when the realized value of X is x. It is *proper* if it encourages honesty in the sense that the expected score $\mathrm{E}_{X\sim P}S(X, Q)$, where X has distribution P, is minimized by the choice $Q = P$.

Traditionally, a scoring rule has been termed *local* if it depends on the density function $q(\cdot)$ of Q only through its value, $q(x)$, at x. With this definition, any proper local scoring rule is equivalent to the *log score*, $S(x, Q) = -\ln q(x)$. However, we can weaken the locality condition by allowing further dependence on a finite number m of derivatives of $q(\cdot)$ at x, and this introduces many further possibilities. We term m the *order* of the rule.

In this paper we describe a large class of such *order-m proper local scoring rules* for densities on the real line. These turn out to depend on the density $q(\cdot)$ in a way that is insensitive to a multiplicative constant, and hence can be computed without knowledge of the normalizing constant of q.

Hyvärinen (2005) proposed a method for approximating a distribution P on $\mathcal{X} = \mathbb{R}^k$ by a distribution Q in a specified family $\mathcal{P}$ of distributions by minimizing $d(P, Q)$ over $Q \in \mathcal{P}$, where

$$d(P, Q) = \frac{1}{2}\int \mathrm{d}x\, p(x)|\nabla \ln p(x) - \nabla \ln q(x)|^2 \tag{1}$$

with ∇ denoting gradient. Since q enters this expression only through $\nabla \ln q$, it is clear that the minimization only requires knowledge of q up to a multiplicative factor. Using integration by parts, Hyvärinen (2005) further showed that minimization

Received January 2011; revised January 2012.

[1]Supported by EPSRC Statistics Mobility Fellowship EP/E009670.

MSC2010 subject classifications. Primary 62C99; secondary 62A99.

Key words and phrases. Bregman score, concavity, divergence, entropy, Euler–Lagrange equation, homogeneity, integration by parts, local function, score matching, variational methods.

of the divergence $d(P, Q)$ in (1) is equivalent to mimimizing

$$S(P, Q) = \mathrm{E}_P\{\Delta \ln q(X) + \tfrac{1}{2}|\nabla \ln q(X)|^2\}$$

[where Δ denotes the Laplacian operator $\sum_{i=1}^k \partial^2/(\partial x_i)^2$], which is a scoring rule of the type discussed in this paper: see Section 2.5 below.

The plan of the paper is as follows. In Section 2 we introduce proper scoring rules, with some examples and applications. Section 3 formalizes the notion of a local function, its representations and derivatives. In Section 4 we apply integration by parts and the calculus of variations to develop a "key equation," which is further investigated in Section 5 through an analysis of fundamental differential operators associated with local functions. Section 6 describes the solutions to the key equation, which we term "key local scoring rules," in terms of a homogeneous function ϕ. In Section 7 we point out that distinct choices of ϕ can generate the same scoring rule, and consider some implications; in particular, we show that key m-local scoring rules exist for any even order m, but for no odd order. Section 9 examines when this construction does indeed yield a proper local scoring rule, concavity of ϕ being crucial. Section 10 devotes further attention to boundary terms arising in the integration by parts. In Section 11 we study how the problem and its solution transform under an invertible mapping of the sample space, and develop an invariant formulation.

1.1. *Related work.* In this paper we are concerned with characterizing m-local proper scoring rules, for all orders m. Since there are no such rules of order 1, order 2 scoring rules constitute the simplest nontrivial case, and as such are likely to be the most useful in practice. In a companion paper to this one, Ehm and Gneiting (2012) conduct a deep investigation of order 2 proper local rules, using an elegant construction complementary to ours. They also describe a general class of densities for which the boundary terms vanish.

The present paper confines attention to absolutely continuous distributions on the real line. The notion of local scoring rule has an interesting analogue for a discrete sample space equipped with a given neighborhood structure. The theory for that case is developed in an accompanying paper [Dawid, Lauritzen and Parry (2012)]; it exhibits both close parallels with, and important differences from, the continuous case considered here.

2. Scoring rules. Suppose You are required to express Your uncertainty about an unobserved quantity $X \in \mathcal{X}$ by quoting a distribution Q over $\mathcal{X}$, after which Nature will reveal the value x of X. A *scoring rule* or *score* S [Dawid (1986)] is a special kind of loss function, intended to measure the quality of your quote Q in the light of the realized outcome x: $S(x, Q)$ is a real number interpreted as the loss You will suffer in this case. The principles of Bayesian decision theory [Savage (1954)] now enjoin You to minimize Your expected loss. If Your actual beliefs about X are described by a probability distribution P, You should thus quote that Q that minimizes $S(P, Q) := \mathrm{E}_{X\sim P} S(X, Q)$. The scoring rule S is termed *proper* (relative to a class $\mathcal{P}$ of distributions over $\mathcal{X}$) when, for any fixed $P \in \mathcal{P}$, the minimum over $Q \in \mathcal{P}$ is achieved at $Q = P$; it is *strictly proper* when, further, this minimum is unique. Thus, under a proper scoring rule, honesty is the best policy.

Associated with any proper scoring rule S are a (*generalized*) *entropy function* $H(P) := S(P, P)$ and a *divergence function* $d(P, Q) := S(P, Q) - H(P)$. Under suitable technical conditions, proper scoring rules and their associated entropy functions and divergence functions enjoy certain properties that serve to characterize such "coherent" constructions [Dawid (1998)]: $S(P, Q)$ is affine in P and is minimized in Q at $Q = P$; $H(P)$ is concave in P; $d(P, Q) - d(P, Q_0)$ is affine in P, and $d(P, Q) \geq 0$, with equality achieved at $Q = P$.

If two scoring rules differ by a function of x only, then they will yield the identical divergence function. In this case we will term them *equivalent* [note that this is a more specialized usage than that of Dawid (1998)].

A fairly arbitrary statistical decision problem can be reduced to one based on a proper scoring rule. Let $L : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ be a loss function, defined for outcome space $\mathcal{X}$ and action space $\mathcal{A}$. Letting $\mathcal{P}$ be a class of distributions over $\mathcal{X}$ such that $L(P, a) := \mathrm{E}_{X\sim P} L(X, a)$ exists for all $a \in \mathcal{A}$ and $P \in \mathcal{P}$, define, for $P, Q \in \mathcal{P}$ and $x \in \mathcal{X}$,

$$S(x, Q) := L(x, a_Q), \tag{2}$$

where $a_P := \arg\inf_{a\in\mathcal{A}} L(P, a)$ is a Bayes act with respect to P (supposed to exist, and selected arbitrarily if nonunique). Then S is readily seen to be a proper scoring rule, and the associated entropy function is just the Bayes loss: $H(P) = \inf_{a\in\mathcal{A}} L(P, a)$.

In this paper we focus attention on the case that $\mathcal{X}$ is an interval on the real line and any $Q \in \mathcal{P}$ has a density $q(\cdot)$ with respect to Lebesgue measure on $\mathcal{X}$. We may then define $S(x, Q)$ in terms of q. However, since q is only defined almost everywhere we must take care that any manipulations performed either involve a preferred version of q, or yield the same answer when q is changed on a null set. This will always be the case in this paper.

2.1. *Bregman scoring rule.* Since any decision problem generates a proper scoring rule there is a very great number of these. Certain forms are of special interest or simplicity. Here we describe one important class of such rules for the case that every $Q \in \mathcal{P}$ has a density function $q(\cdot)$ with respect to a dominating measure μ over $\mathcal{X}$.

Let $\phi : \mathbb{R}^+ \to \mathbb{R}$ be concave and differentiable. The associated (*separable*) *Bregman scoring rule* is defined by

$$S(x, Q) := \phi'\{q(x)\} + \int \mathrm{d}\mu(y)\,[\phi\{q(y)\} - q(y)\phi'\{q(y)\}]. \tag{3}$$

It can be shown that these are the only proper scoring rules having the form $S(x, Q) = \xi\{q(x)\} - k(Q)$ [Dawid (2007)].

Taking expectations, we obtain

$$S(P, Q) = \int d\mu(x)\, [\{p(x) - q(x)\}\phi'\{q(x)\} + \phi\{q(x)\}].$$

It follows that $H(P) = \int d\mu(x)\, \phi\{p(x)\}$ and so, assuming $H(P)$ is finite, the corresponding (separable) *Bregman divergence* [Bregman (1967), Csiszár (1991)]—also termed *U-divergence* [Eguchi (2008)]—is

$$d(P, Q) = \int d\mu(x)\, \bigl([\phi\{q(x)\} + \{p(x) - q(x)\}\phi'\{q(x)\}] - \phi\{p(x)\}\bigr). \tag{4}$$

The integrand is nonnegative by concavity of ϕ. Therefore, the separable Bregman scoring rule is a proper scoring rule, and strictly proper if ϕ is strictly concave.

2.2. *Extended Bregman score*. A straightforward generalization of the above Bregman construction is obtained on replacing $\phi : \mathbb{R}^+ \to \mathbb{R}$ throughout by $\phi : \mathcal{X} \times \mathbb{R}^+ \to \mathbb{R}$, such that, for each $x \in \mathcal{X}$, $\phi(x, \cdot) : \mathbb{R}^+ \to \mathbb{R}$ is concave. Such *extended Bregman* rules are the only proper scoring rules of the form $S(x, Q) = \xi\{x, q(x)\} - k(Q)$ [Dawid (2007)].

2.3. *Log score*. For $\phi(s) \equiv -s \ln s$ we obtain the *logarithmic scoring rule*, or *log score*, defined by

$$S(x, Q) = -\ln q(x).$$

This is essentially the only scoring rule of the form $S(x, Q) = \xi\{x, q(x)\}$ [Bernardo (1979), Dawid (2007)]. For this case we obtain

$$H(P) = -\int d\mu(x)\, p(x) \ln p(x),$$

the *Shannon entropy*, and

$$d(P, Q) = \int d\mu(x)\, p(x) \ln \frac{p(x)}{q(x)}, \tag{5}$$

the *Kullback–Leibler divergence*.

2.4. *Parameter estimation*. Let $\mathcal{Q} = \{Q_\theta\} \subseteq \mathcal{P}$ be a smooth parametric family of distributions. Given data $(x_1, \dots, x_n)$ in $\mathcal{X}$ with empirical distribution $\widehat{P}$, one way to estimate θ is by minimizing some divergence criterion: $\widehat{\theta} := \arg\min_\theta d(\widehat{P}, Q_\theta)$. When the divergence function is derived from a scoring rule, this is equivalent to minimizing the total empirical score:

$$\widehat{\theta} = \arg\min_\theta \sum_{i=1}^{n} S(x_i, Q_\theta), \tag{6}$$

in which form it remains meaningful even if $\widehat{P} \notin \mathcal{P}$, when $d(\widehat{P}, Q_\theta)$ is undefined. The corresponding estimating equation is

$$\sum_{i=1}^{n} \sigma(x_i, \theta) = 0 \tag{7}$$

with $\sigma(x, \theta) := \partial S(x, Q_\theta)/\partial\theta$. For a proper scoring rule it is straightforward to show that the estimating equation (7) is unbiased [Dawid and Lauritzen (2005)] and, as a result, $\widehat{\theta}$ is typically consistent, though not necessarily efficient; it may also display some degree of robustness. Equation (7) delivers an *M-estimator* [Huber (1981), Hampel et al. (1986)]. Statistical properties of the estimator are considered by Eguchi (2008) for the special case of minimum Bregman (U-) divergence estimation, and readily extend to more general cases.

2.5. *Hyvärinen scoring rule.* Hyvärinen (2005) showed that minimization of the divergence $d(P, Q)$ in (1) is equivalent to minimizing the empirical score for the scoring rule

$$S(x, Q) = \Delta \ln q(x) + \tfrac{1}{2}|\nabla \ln q(x)|^2. \tag{8}$$

This is valid in the case where $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{P}$ consists of distributions P whose Lebesgue density $p(\cdot)$ is a twice continuously differentiable function of x satisfying $\nabla \ln p \to 0$ as $|x| \to \infty$. For $k = 1$ we get

$$S(x, Q) = \frac{q''(x)}{q(x)} - \frac{1}{2}\left\{\frac{q'(x)}{q(x)}\right\}^2. \tag{9}$$

Dawid and Lauritzen (2005) showed that, with some reinterpretation, the formula (8) defines a proper scoring rule in the more general case of an outcome space $\mathcal{X}$ that is a Riemannian manifold. Now $q(\cdot)$ denotes the *natural density* $\mathrm{d}Q/\mathrm{d}\mu$ of Q with respect to the associated volume measure μ on $\mathcal{X}$; ∇ denotes *natural gradient*; Δ is the *Laplace–Beltrami operator*; and $|u|^2 = \langle u, u\rangle$ is the squared norm defined by the metric tensor. We impose the restriction $P, Q \in \mathcal{P}$, where $P \in \mathcal{P}$ if $\nabla \ln p(x) \to 0$ as x approaches the boundary of $\mathcal{X}$.

On applying Stokes's theorem (again essentially integration by parts) and noting that boundary terms vanish, we can express the expected score as

$$S(P, Q) = \frac{1}{2}\int \mathrm{d}\mu(x)\, p(x)\langle \nabla \ln q(x) - 2\nabla \ln p(x), \nabla \ln q(x)\rangle.$$

The entropy is thus $H(P) = -\frac{1}{2}\int \mathrm{d}\mu(x)\, p(x)|\nabla \ln p(x)|^2$ and so the associated divergence is essentially that used by Hyvärinen:

$$d(P, Q) = \frac{1}{2}\int \mathrm{d}\mu(x)\, p(x)|\nabla \ln p(x) - \nabla \ln q(x)|^2, \tag{10}$$

which is nonnegative and vanishes only when $Q = P$. It follows that the scoring rule is strictly proper.

Although this scoring rule is not local in the strict sense, it depends on (x, Q) only through the first and second derivatives of the density function $q(\cdot)$ at the point x; it is *local of order* 2, or 2-*local*, as defined below in Section 3.

Note that one does not need to know the volume measure μ to calculate the divergence; formula (10) for $d(P, Q)$ yields the same result if we take μ to be any fixed underlying measure, and interpret p and q as densities with respect to this.

2.6. *Homogeneity.* An interesting and practically valuable property of the generalized Hyvärinen scoring rule is that $S(x, Q)$ given by (8) is *homogeneous* in the density function $q(\cdot)$: it is formally unchanged if $q(\cdot)$ is multiplied by a positive constant, and so can be computed even if we only know the density function up to a scale factor. In particular, use of the estimating equation (7) does not require knowledge of the normalizing constant (which is often hard to obtain) for densities in $\mathcal{Q}$.

EXAMPLE 2.1. Consider the natural exponential family:

$$q(x|\theta) = Z(\theta)^{-1} \exp\{a(x) + \theta x\}.$$

Using the scoring rule (9) we obtain $S(x, Q_\theta) = a''(x) + \frac{1}{2}\{a'(x) + \theta\}^2$, so that $\sigma(x, \theta) = a'(x) + \theta$, and (7) delivers the unbiased estimator

$$\widehat{\theta} = -\sum_{i=1}^{n} a'(X_i)/n,$$

which can be computed without knowledge of $Z(\theta)$. See also Section 4 of Hyvärinen (2007), where exponential families are discussed.

Alternatively, we can work directly with the sufficient statistic $T := \sum_{i=1}^{n} X_i$, which has density of the form

$$q_T(t|\theta) = Z(\theta)^{-n} \exp\{\alpha_n(t) + \theta t\}.$$

Applying the above method to q_T leads to the unbiased estimator $\widetilde{\theta} = -\alpha_n'(T)$. This is the *maximum plausibility estimator* of θ [Barndorff-Nielsen (1976)]. Basing the estimate on the sufficient statistic is more satisfying and better behaved from a principled point of view, but does require computation of the function $\alpha_n(t)$, which involves an n-fold convolution of $e^{a(x)}$.

As an application, suppose Q_θ is obtained from the normal distribution $N(\theta, 1)$ by retaining its outcome x with probability $k(x)$. We assume that $k(x)$ is everywhere positive and twice differentiable. The density is thus

$$q(x|\theta) = \frac{k(x) \exp -\{(x-\theta)^2/2\}}{\int k(y) \exp -\{(y-\theta)^2/2\}\, dy}.$$

Because of the complex dependence of the denominator on θ, the maximum likelihood estimate typically cannot be expressed in closed form. However, using scoring rule (9) yields the explicit unbiased estimator

$$\widehat{\theta} = \sum_{i=1}^{n} \{x_i - \kappa'(x_i)\}/n$$

with $\kappa(x) := \ln k(x)$.

The homogeneity property will be a feature of all the new proper local scoring rules we introduce here: see Section 6.

3. Local scoring rules. We observed in Section 2.3 that the log score $S(x, Q) = -\ln q(x)$ is essentially the only proper scoring rule that is *local*, that is, involves the density function $q(\cdot)$ of Q only through its value, $q(x)$, at the actually realized value x of X.

We can, however, weaken the locality requirement, for example, by allowing S to depend on the values of $q(\cdot)$ in an infinitesimal neighborhood of x. In this paper we describe a class of scoring rules that depend on the function $q(\cdot)$ only through its value and the values of a finite number m of its derivatives at the point x—a property we will refer to as *locality of order* m, or m*-locality*.

We confine ourselves to the case that $\mathcal{X}$ is an open interval in $\mathbb{R}$, possibly infinite or semi-infinite, and $\mathcal{P}$ is a class of distributions Q over $\mathcal{X}$ having strictly positive Lebesgue density $q(\cdot)$ that is m-times continuously differentiable.

3.1. *Local functions and scoring rules*. To study the properties of local scoring rules we need a formal definition of a local function.

DEFINITION 3.1. A function $F : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ is said to be *local of order* m, or m*-local*, if it can be expressed in the form

$$F(x, Q) = f\{x, q(x), q'(x), q''(x), \dots, q^{(m)}(x)\},$$

where $f : \mathcal{X} \times \mathbf{Q}_m \to \mathbb{R}$, with $\mathbf{Q}_m := \mathbb{R}^+ \times \mathbb{R}^m$, is a real-valued infinitely differentiable function, $q(\cdot)$ is the density function of Q, and a prime ($'$) denotes differentiation with respect to x. It is *local* if it is local of some finite order.

We shall refer to such a function f as a q*-function*, and say it is of order m. When we do not need to specify the order m of a q-function f we may write $f(x, \mathbf{q})$ ($x \in \mathcal{X}$, $\mathbf{q} \in \mathbf{Q}$), understanding $\mathbf{Q} = \mathbf{Q}_m$, $\mathbf{q} = (q_0, \dots, q_m)$.

A scoring rule $S(x, Q)$ is m*-local* if

$$S(x, Q) = s\{x, q(x), q'(x), q''(x), \dots, q^{(m)}(x)\}, \tag{11}$$

where s is a q-function as above, so that it depends on the quoted distribution Q for X only through the value and derivatives up to order m of the density $q(\cdot)$ of Q, evaluated at the observed value x of X. The function s is the *score function* of S.

3.2. *Differentiation of local functions.* For a local scoring rule S given by (11) we write

$$S_{[j]}(x, Q) := s_{[j]}\{x, q(x), q'(x), q''(x), \dots, q^{(m)}(x)\},$$

where $s_{[j]} := \partial s/\partial q_j$, and similarly

$$S_{[x]}(x, Q) := s_{[x]}\{x, q(x), \dots, q^{(m)}(x)\},$$

where $s_{[x]} := \partial s/\partial x$. Then if $\mathrm{d}S/\mathrm{d}x$ denotes the derivative of $S(x, Q)$ with respect to x for fixed Q, we have

$$\frac{\mathrm{d}S}{\mathrm{d}x} = S_{[x]}(x, Q) + \sum_{j\geq 0} q^{(j+1)}(x) S_{[j]}(x, Q). \tag{12}$$

For S of order m, the series in (12) terminates at $j = m$.

Motivated by (12), we introduce a linear differential operator D acting on q-functions by

$$D := \frac{\partial}{\partial x} + \sum_{j\geq 0} q_{j+1} \frac{\partial}{\partial q_j}. \tag{13}$$

For f of order m, the series for Df obtained from (13) terminates at $j = m$, and Df is then of order $m + 1$.

The operator D thus represents the total derivative of the local function for fixed Q:

$$\frac{\mathrm{d}S}{\mathrm{d}x} = (Ds)\{x, q(x), \dots, q^{(m+1)}(x)\},$$

where s is the score function of S.

In the light of the interpretation of D as $\mathrm{d}/\mathrm{d}x$, the following result is unsurprising:

LEMMA 3.1. *For a q-function f, $Df = 0$ if and only if f is constant.*

PROOF. "If" is trivial. For "only if," suppose f is of order $\leq m$. The only term in Df involving q_{m+1} is $q_{m+1} f_{[m]}$, so that $Df = 0 \Rightarrow f_{[m]} = 0$, whence f is of order at most $m - 1$. Repeating this argument, f must be of order 0, that is, of the form $f(x)$. Then $0 = Df = f'(x)$, so finally f must be a constant. $\square$

4. Variational analysis. We are interested in constructing proper local scoring rules. Ideally we would develop sufficient conditions on the score function s and the family $\mathcal{P}$ to ensure that, for any $P \in \mathcal{P}$, $S(P, Q) = \int \mathrm{d}x\, p(x)s\{x, q(x), q'(x), q''(x), \dots, q^{(m)}(x)\}$ is minimized, over $Q \in \mathcal{P}$, at $Q = P$. Initially, however, we shall merely develop, in a somewhat heuristic fashion, conditions sufficient to ensure that, for all $P \in \mathcal{P}$, $Q = P$ will be a *stationary point* of $S(P, Q)$—a property we shall describe by saying that S is a *weakly proper* scoring rule. Given any S satisfying these conditions, further attention will be required to check whether or not it is in fact proper; this will be taken up in Section 9 below.

To address this problem we adopt the methods of variational calculus [Troutman (1983), van Brunt (2004)]. Suppose that, at $Q = P$, $S(P, Q)$ is stationary under an arbitrary infinitesimal variation $\delta q(\cdot)$ of $q(\cdot)$, subject to the requirement that $q(\cdot) + \delta q(\cdot)$ be a probability density. That is,

$$\begin{aligned}\delta\left\{\int \mathrm{d}x\, p(x) s\{x, q(x), q'(x), q''(x), \ldots, q^{(m)}(x)\} + \lambda_P \int \mathrm{d}x\, q(x)\right\}\bigg|_{q=p}\\ = 0,\end{aligned} \tag{14}$$

where λ_P is a Lagrange multiplier for the normalization constraint $\int \mathrm{d}x\, q(x) = 1$. The left-hand side of (14), evaluated with $P = Q$, is

$$\int \mathrm{d}x \left\{\sum_{k=0}^{m} \delta q^{(k)}(x) q(x) S_{[k]}(x, Q) + \lambda_Q \delta q(x)\right\}, \tag{15}$$

and this is to vanish for arbitrary infinitesimal $\delta q(\cdot)$ and suitable λ_Q.

We evaluate the integral of the kth term of the sum in (15) using the general formula for repeated integration by parts:

$$\begin{aligned}&(-1)^k \int_-^+ \mathrm{d}x\, F G^{(k)}\\ &\quad = \int_-^+ \mathrm{d}x\, G F^{(k)} - \sum_{r=0}^{k-1} (-1)^{k-1-r} \{G^{(k-1-r)} F^{(r)}\}\big|_-^+,\end{aligned} \tag{16}$$

where $F^{(k)}$ denotes the kth derivative of F with respect to x. The first term on the right-hand side of (16) is the *integral term*; the remaining terms are *boundary terms*, these being evaluated, if necessary, as limits as we approach, from within, the end-points (denoted by $-$ and $+$) of the interval $\mathcal{X} \subseteq \mathbb{R}$.

Setting $G = q(x) S_{[k]}(x, Q)$, $F = \delta q(x)$, we obtain

$$\begin{aligned}&\int_-^+ \mathrm{d}x\, q(x) S_{[k]}(x, Q) \delta q^{(k)}(x)\\ &\quad = \int_-^+ \mathrm{d}x\, (-1)^k \delta q(x) \frac{\mathrm{d}^k}{\mathrm{d}x^k} \{q(x) S_{[k]}(x, Q)\}\\ &\qquad + \sum_{r=0}^{k-1} (-1)^{k-1-r} \frac{\mathrm{d}^{k-1-r}}{\mathrm{d}x^{k-1-r}} \{q(x) S_{[k]}(x, Q)\} \delta q^{(r)}(x)\big|_-^+.\end{aligned} \tag{17}$$

At this point we restrict consideration to functions δq whose derivatives vanish sufficiently quickly at the end-points that we can suppose the boundary terms in the last line of (17) vanish. Then (15) will vanish for all such $\delta q(\cdot)$ if

$$\sum_{k=0}^{m}(-1)^{k+1}\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left\{q(x)S_{[k]}(x,Q)\right\}\equiv\lambda_Q, \tag{18}$$

that is, the left-hand side of (18) is a constant, independent of x.

Motivated by (18), we introduce the following linear differential operator L on q-functions:

$$L:=\sum_{k\geq 0}(-1)^{k+1}D^k q_0\frac{\partial}{\partial q_k}. \tag{19}$$

Unless overridden by parentheses, operators here and elsewhere associate to the right, so that $Tq_0 f$ means $T(q_0 f)$, that is, we have

$$Lf=\sum_{k\geq 0}(-1)^{k+1}D^k\left(q_0\frac{\partial f}{\partial q_k}\right). \tag{20}$$

For f of order m, the series in (20) terminates at $k=m$, and the order of Lf is at most $2m$.

We can now write (18) as

$$Ls=\lambda_Q, \tag{21}$$

where equality in (21) is required to hold for all $(x,q_0,q_1,\ldots,q_{2m})$ such that $q_j=q^{(j)}(x)$ $(j=0,\ldots,2m)$. In particular, a *sufficient* condition that S be weakly proper is that for some $\lambda\in\mathbb{R}$ we have

$$Ls=\lambda \tag{22}$$

for all $x\in\mathcal{X}$, $\mathbf{q}\in\mathbf{Q}_{2m}$.

So long as $\mathcal{P}$ is sufficiently large, the form (22) will also be necessary for (21) to hold. In particular, suppose we impose the following condition on $\mathcal{P}$:

CONDITION 4.1. Given distinct $x_1,x_2\in\mathcal{X}$, and any $\mathbf{q}_1,\mathbf{q}_2\in\mathbf{Q}_{2m}$, there exists $Q\in\mathcal{P}$ satisfying $q^{(j)}(x_1)=q_{1,j}$, $q^{(j)}(x_2)=q_{2,j}$ $(j=0,\ldots,2m)$.

Take arbitrary $Q_1,Q_2\in\mathcal{P}$, $x_1\neq x_2\in\mathcal{X}$, and set $q_{i,j}:=q_i^{(j)}(x_i)$ $(i=1,2;\ j=0,\ldots,2m)$. Let Q be as given by Condition 4.1. Evaluating (21) at $(x_1,\mathbf{q}_1)$ yields $\lambda_{Q_1}=\lambda_Q$, and similarly $\lambda_{Q_2}=\lambda_Q$. Thus λ_Q cannot depend on Q, and so (22) must hold. Moreover, taking $x_1=x$, $\mathbf{q}_1=\mathbf{q}$, and x_2, $\mathbf{q}_2$ arbitrary, this must hold for any $x\in\mathcal{X}$, $\mathbf{q}\in\mathbf{Q}_{2m}$.

So we henceforth restrict attention to solutions of (22). We note that a particular solution of (22) is given by the *log-score*:

$$s=-\lambda\ln q_0.$$

Since L is a linear operator, the general solution is of the form

$$s = -\lambda \ln q_0 + s_0,$$

where s_0 satisfies the *key equation*:

$$Ls_0 = 0. \tag{23}$$

Because of this we shall confine attention to solutions of the key equation, and shall term any solution of (23) a *key local score function*.

4.1. *Connection to classical calculus of variations.* Because the Lagrange multiplier λ associated with a key local scoring rule s vanishes, setting $q(x) \equiv p(x)$ will in fact deliver a *globally* stationary point [i.e., without imposing the normalization constraint $\int \mathrm{d}x\, q(x) = 1$] of the corresponding expected score

$$\int \mathrm{d}x\; p(x) s\{x, q(x), q'(x), q''(x), \dots, q^{(m)}(x)\}.$$

The classical calculus of variations—see, for example, van Brunt [(2004), equations (2.9), (3.3)]—would (again, ignoring the boundary terms) identify the solution to this unconstrained variational problem in $q(\cdot)$ as solving the *Euler–Lagrange equation*:

$$\Lambda p_0 s(x, q_0, \dots, q_m) = 0, \tag{24}$$

where Λ is the *Lagrange operator*:

$$\Lambda := \sum_{k \geq 0} (-1)^k D^k \frac{\partial}{\partial q_k}. \tag{25}$$

We want the solution of (24) to be $\mathbf{q} = \mathbf{p}$.

Now when evaluated at $\mathbf{q} = \mathbf{p}$, $\Lambda q_0 s = s + \Lambda p_0 s$. So $\mathbf{q} = \mathbf{p}$ should satisfy

$$(I - \Lambda q_0)s = 0, \tag{26}$$

where I is the identity operator.

But

$$\frac{\partial}{\partial q_0}(q_j f) = \begin{cases} q_j \dfrac{\partial f}{\partial q_0}, & (j > 0), \\[2ex] f + q_0 \dfrac{\partial f}{\partial q_0}, & (j = 0), \end{cases}$$

so we have

$$\begin{aligned} L &= I + \sum_{k \geq 0} (-1)^{k+1} D^k \frac{\partial}{\partial q_k} q_0 \circ \\ &= I - \Lambda q_0 \circ. \end{aligned} \tag{27}$$

(Here and throughout, for g a q-function, $g\circ$ denotes the multiplication operator $f \mapsto gf$, the optional symbol $\circ$ being attached, where required, to avoid confusion with the q-function g itself.)

Hence (26) becomes $Ls = 0$, so recovering the key equation (23).

5. Properties of differential operators. For a further study of the key equation and its properties we shall have a detailed look at the differential operators introduced earlier, together with some new ones. We recall

$$\begin{aligned} D &= \frac{\partial}{\partial x} + \sum_{j \geq 0} q_{j+1} \frac{\partial}{\partial q_j}, \\ L &= \sum_{k \geq 0} (-1)^{k+1} D^k q_0 \frac{\partial}{\partial q_k}, \\ \Lambda &= \sum_{k \geq 0} (-1)^k D^k \frac{\partial}{\partial q_k} \\ &= (I - L) q_0^{-1} \circ. \end{aligned}$$

LEMMA 5.1. *The Lagrange operator Λ annihilates the total derivative operator D*:

$$\Lambda D = 0. \tag{28}$$

PROOF. Using $(\partial/\partial q_k) D = D(\partial/\partial q_k) + (\partial/\partial q_{k-1})$ we have

$$\begin{aligned} \Lambda D &= \sum_{k \geq 0} (-1)^k D^{k+1} \frac{\partial}{\partial q_k} + \sum_{k \geq 0} (-1)^k D^k \frac{\partial}{\partial q_{k-1}} \\ &= 0. \end{aligned}$$ □

We now introduce the *Euler operator*:

$$E := \sum_{j \geq 0} q_j \frac{\partial}{\partial q_j}. \tag{29}$$

LEMMA 5.2. *The Euler operator E commutes with D and with L, while*

$$\Lambda E = E \Lambda + \Lambda. \tag{30}$$

PROOF. From the easily verified relations

$$\begin{aligned} E q_j \circ &= q_j \circ + q_j E, \\ (\partial/\partial q_k) E &= E(\partial/\partial q_k) + (\partial/\partial q_k), \end{aligned} \tag{31}$$

it readily follows that E commutes with $q_j(\partial/\partial q_k)$. Since clearly E commutes with $\partial/\partial x$, E thus commutes with D, and consequently with any power of D. From (19), we now see that E commutes with L.

Now (27) gives that E commutes with $\Lambda q_0\circ$, and thus, noting from (31) that $q_0 E q_0^{-1}\circ = E - I$, we have

$$E\Lambda = E\Lambda q_0 q_0^{-1}\circ = \Lambda q_0 E q_0^{-1}\circ = \Lambda(E - I) = \Lambda E - \Lambda,$$

which yields (30). $\square$

THEOREM 5.3. *We have that* $\Lambda E = \Lambda q_0 \Lambda$.

PROOF. Using $q_k D = D q_k\circ - q_{k+1}\circ$, we can readily show by induction that, for $k \geq 0$,

$$q_0 D^k = D\left\{\sum_{j=0}^{k-1}(-1)^j q_j D^{k-1-j}\right\} + (-1)^k q_k\circ.$$

It now follows from (28) that $\Lambda q_0 D^k = (-1)^k \Lambda q_k\circ$. Applying this term-by-term to (25) we obtain $\Lambda q_0 \Lambda = \sum_k \Lambda q_k(\partial/\partial q_k) = \Lambda E$. $\square$

For later purposes we introduce, for any integer r,

$$B_r := \sum_{k\geq r+1}(-1)^{k-1-r} D^{k-1-r}\frac{\partial}{\partial q_k} \tag{32}$$

with the understanding $\partial/\partial q_k = 0$ if $k \leq 0$ (in particular, $B_{-1} = \Lambda$). We further define

$$C := \sum_{r\geq 0} q_r B_r. \tag{33}$$

LEMMA 5.4. *It holds that*

$$DB_r = (\partial/\partial q_r) - B_{r-1}, \tag{34}$$

$$B_r D = (\partial/\partial q_r). \tag{35}$$

PROOF. Equation (34) follows easily from the definition (32), while (35) can be proved in the same way as Lemma 5.1. $\square$

THEOREM 5.5. *We have*

$$CD = E, \tag{36}$$

$$DC = E - q_0\Lambda. \tag{37}$$

PROOF. Equation (36) follows directly from (33) and (35). From (34), $Dq_r B_r = q_{r+1}B_r - q_r B_{r-1} + q_r(\partial/\partial q_r)$, and thus $DC = \sum_{r\geq 0} q_r(\partial/\partial q_r) - q_0 B_{-1} = E - q_0\Lambda$. $\square$

6. Homogeneous scoring rules. We shall see that all key local score functions, that is, solutions to the key equation $Ls = 0$, are homogeneous in the sense that changing $\mathbf{q}$ by a multiplicative factor does not change the value of s; hence the associated scoring rule $S(x, Q)$ only involves the density q up to a constant factor. We shall formalize and show this below.

DEFINITION 6.1. A q-function f is said to be *homogeneous of degree h*, or *h-homogeneous*, if, for any $\lambda > 0$, $f(x, \lambda\mathbf{q}) \equiv \lambda^h f(x, \mathbf{q})$.

With E defined by (29), Euler's homogeneous function theorem implies that a q-function f is h-homogeneous if and only

$$Ef = hf. \tag{38}$$

The partial derivatives $f_{[j]} = \partial f/\partial q_j$ of an h-homogeneous function are homogeneous of order $h - 1$, while $f_{[x]} = \partial f/\partial x$ is homogeneous of order h. It follows that, if f is h-homogeneous, then so is Df.

In this work we shall only need to deal with homogeneity of degree 0, where $f(x, \lambda\mathbf{q}) \equiv f(x, \mathbf{q})$, and of degree 1, where $f(x, \lambda\mathbf{q}) \equiv \lambda f(x, \mathbf{q})$. Clearly, f is 0-homogeneous if and only if $q_0 f$ is 1-homogeneous.

A scoring rule S will be called *homogeneous* if its score function s is 0-homogeneous. As already noted, the logarithmic score is 0-local, but is not homogeneous. The Hyvärinen scoring rule and its generalizations, as described in Section 2.5, are 2-local, and are homogeneous.

We can now easily show that key local score functions are homogeneous:

THEOREM 6.1. *If $Lf = 0$, then f is 0-homogeneous.*

PROOF. In this case $f = (I - L)f = \Lambda q_0 f$ and thus from (30) and Theorem 5.3 we get

$$Ef = E\Lambda q_0 f = \Lambda E q_0 f - \Lambda q_0 f = \Lambda q_0(\Lambda q_0 f) - \Lambda q_0 f = \Lambda q_0 f - \Lambda q_0 f = 0$$

as required. □

We can further show that, if we consider the restriction of the operator L to 0-homogeneous functions, it acts as a *projection operator*; $I - L$ is then the complementary projection. This is a consequence of the fact that L is idempotent when restricted to 0-homogeneous functions:

THEOREM 6.2. *If f is 0-homogeneous, then so is Lf, and $L^2 f = Lf$.*

PROOF. Since E commutes with L, if f is 0-homogeneous we have $ELf = LEf = 0$, so Lf is 0-homogeneous as well. If f is 0-homogeneous, $q_0 f$ is 1-homogeneous and thus by (38) and Theorem 5.3

$$\Lambda q_0 f = \Lambda E q_0 f = \Lambda q_0 \Lambda q_0 f,$$

so $I - L = \Lambda q_0 \circ$ is idempotent when restricted to 0-homogeneous functions, whence so is L. $\square$

Elaborating the consequences of these results we get:

COROLLARY 6.3. *We have that*:

(i) $Lf = 0$ *if and only if* $f = (I - L)g$ *for some* 0-*homogeneous* g; *equivalently*, $f = \Lambda\phi$ *for some* 1-*homogeneous* ϕ;

(ii) *if* f *is* 0-*homogeneous*, *then* $(I - L)f = 0$ *if and only if* $f = Lg$ *for some* 0-*homogeneous* g.

PROOF. If $Lf = 0$, then f is 0-homogeneous by Theorem 6.1. The other properties are easy consequences of the fact that L and $I - L$ are complementary projections in the space of 0-homogeneous functions. $\square$

Collecting everything, we have the following main result:

THEOREM 6.4. *A* q-*function* s *is a key local score function if and only if any one* (*and then all*) *of the following conditions holds*:

(i) *The function* s *satisfies the key equation* $Ls = 0$, *where the operator* L *is given by* (19).

(ii) *We can express* $s = (I - L)g$ *where* g *is a* 0-*homogeneous* q-*function*.

(iii) *We can express* $s = \Lambda\phi$ *where* ϕ *is a* 1-*homogeneous* q-*function and the operator* Λ *is given by* (25).

Moreover, s *is then* 0-*homogeneous*.

When (ii) above holds, we say that s is *derived from* g; when (iii) holds, we say that s is *generated by* ϕ. The key local score function generated by a 1-homogeneous q-function ϕ of order t is thus

$$s(x, \mathbf{q}) = \sum_{k=0}^{t} (-1)^k D^k \phi_{[k]}(x, \mathbf{q}).$$

The only term in s that involves q_{2t} is $(-1)^t \phi_{[tt]} q_{2t}$. In particular, if $\phi_{[tt]} \neq 0$, s is of exact order $2t$. Hence we have demonstrated the existence of key local scoring rules of all positive even orders.

The key local scoring rule S generated by ϕ is then

$$S(x, Q) = \sum_{k=0}^{t} (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} \phi_{[k]}\{x, q(x), q'(x), \dots, q^{(t)}(x)\}. \tag{39}$$

For the case $t = 1$ we obtain a second-order rule:

$$S(x, Q) = \phi_{[0]}\{x, q(x), q'(x)\} - \frac{\mathrm{d}}{\mathrm{d}x}\phi_{[1]}\{x, q(x), q'(x)\},$$

where $\phi(x, q_0, q_1)$ is 1-homogeneous.

The Hyvärinen scoring rule (9) is generated in this way by $\phi = -\frac{1}{2}q_1^2/q_0$. More generally, choosing $\phi = -q_1^k/q_0^{k-1}$ $(k \geq 1)$ yields

$$S(x, Q) = (k-1)(y_1^k + k y_1^{k-2} y_2), \tag{40}$$

where $y_i := (\mathrm{d}^i/\mathrm{d}x^i)\ln q(x)$. We can express a general 1-homogeneous x-independent q-function of order 1 as a power series:

$$\phi(q_0, q_1) = q_0 \sum_{k\geq 1} a_k (q_1/q_0)^k. \tag{41}$$

Now combining the rules (40) arising from the individual terms in (41), we obtain the series form of a general x-independent second-order scoring rule described by Ehm and Gneiting (2010).

7. Gauge transformation. The map $\phi \mapsto s = \Lambda\phi$ in Theorem 6.4(iii) is many-to-one: two 1-homogeneous functions ϕ_1 and ϕ_2 will generate the identical score function $s = \Lambda\phi_1 = \Lambda\phi_2$ if and only if $\Lambda(\phi_2 - \phi_1) = 0$. And this will hold if and only if $\phi_2 - \phi_1$ has the total derivative form $D\psi$:

LEMMA 7.1. *Suppose ϕ is 1-homogeneous. Then $\Lambda\phi = 0$ if and only if ϕ has the form $D\psi$.*

PROOF. If $\phi = D\psi$, then $\Lambda\phi = 0$ by (28). Conversely, suppose ϕ is 1-homogeneous and $\Lambda\phi = 0$. Then $q_0^{-1}\phi$ is 0-homogeneous and $(I - L)q_0^{-1}\phi = 0$, so by Corollary 6.3(ii) there exists 0-homogeneous g such that $q_0^{-1}\phi = Lg$. Now take $\psi = Cq_0 g$, with C given by (33). Then, using (37), $D\psi = (E - q_0\Lambda)q_0 g = (I - q_0\Lambda)q_0 g$, since $Eq_0 g = q_0 g$ because $q_0 g$ is 1-homogeneous; and this is $q_0(I - \Lambda q_0 \circ)g = q_0 L g = \phi$. $\square$

Borrowing terminology from physics, we term a transformation of the form $\phi \to \phi + D\psi$ a *gauge transformation*; the invariance of s under such a transformation of ϕ is *gauge invariance*. The choice of a particular function ϕ, out of the equivalence class of functions differing only by a total derivative $D\psi$ and thus generating the same scoring rule, is a *gauge choice*.

Clearly if $\phi_2 - \phi_1 = D\psi$ and both ϕ_1 and ϕ_2 are 1-homogeneous, then $D\psi$ must be 1-homogeneous. This will be so if ψ is itself 1-homogeneous. The converse also essentially holds:

LEMMA 7.2. *Suppose $D\psi$ is* 1*-homogeneous. Then, for some constant a, $\psi + a$ is* 1*-homogeneous.*

PROOF. We have $ED\psi = D\psi$. Since by Lemma 5.2 D commutes with E, $D(E\psi - \psi) = 0$. Thus by Lemma 3.1 $E\psi - \psi$ is a constant, a say. Then $E(\psi + a) = E\psi = \psi + a$, so $\psi + a$ is 1-homogeneous. □

Since the addition of a constant has no consequences for the analysis, we henceforth call a transformation $\phi \to \phi + \kappa$ a gauge transformation if and only if κ has the form $D\psi$ with ψ 1-homogeneous.

7.1. *Standard gauge choice.* For any key local score function s we note that

$$\phi = q_0 s \tag{42}$$

satisfies

$$\Lambda\phi = \Lambda q_0 s = (I - L)s = s$$

and hence $\phi = q_0 s$ is a valid gauge choice for s. We call (42) the *standard gauge choice*.

7.2. *Equivalence.* Suppose s is generated by ϕ, and let $\phi^* = \phi + \chi$ with $\chi = a(x)q_0$. This is not a gauge transformation if $a \not\equiv 0$, but the score function it generates, $s^* = s + a(x)$, is equivalent to s—which we describe by saying ϕ^* and ϕ are *equivalent*. Conversely, if $\phi^\dagger$ generates $s + a(x)$ it must be a gauge transformation of ϕ^*, and hence of the form $\phi + a(x)q_0 + D\psi$—this form thus being necessary and sufficient for equivalence. We note in particular that $\phi^\dagger$ of the form $\phi + \sum_{k\geq 0} a_k(x)q_k$ is equivalent to ϕ, since it generates $s^\dagger = s + \sum_{k\geq 0}(-1)^k a_k^{(k)}(x)$.

7.3. *Nonexistence of odd-order key local scores.* In Section 6 we established the existence of key local score functions of all positive even orders. Here we show that no key local score function can be of odd order.

Take $s = \Lambda\phi$ as in Theorem 6.4(iii), and suppose s has odd order. If ϕ is of order t, the order of s is at most $2t$; since it is odd, it must be strictly less than $2t$. Again, the only term in s that could possibly involve q_{2t} is $(-1)^t \phi_{[tt]} q_{2t}$, whence $\phi_{[tt]} = 0$. Hence $\phi_{[t]}$, which must be 0-homogeneous, has the form $A(x, q_0, \ldots, q_{t-1})$.

Now define

$$\psi(x, q_0, \ldots, q_{t-1}) := -\int_0^{q_{t-1}} A(x, q_0, \ldots, q_{t-2}, z)\,\mathrm{d}z \tag{43}$$

[for the case $t = 1$ the integrand on the right-hand side is $A(x, z)$]. It is easy to see that ψ is 1-homogeneous, and

$$\phi_{[t]} + \psi_{[t-1]} = 0. \tag{44}$$

Let $\phi^* = \phi + D\psi$, which is of order at most t. Since this is a gauge transformation, ϕ^* generates the same scoring rule s as ϕ does. But from (44), $\phi^*_{[t]} = \phi_{[t]} + \psi_{[t-1]} = 0$, so that ϕ^* is in fact of order at most $t-1$, whence s is of order at most $2t-2$. We can now repeat the argument, stepping down t by 1 each time, until we reach a contradiction.

7.4. *Second-order rule.* A similar argument to the above shows that, for any key local scoring rule of exact even order $2t$, there exists a gauge choice of exact order t.

A second-order rule can thus always be generated by a 1-homogeneous ϕ of order 1. However, a change of gauge may increase the order of the generating function—for example, the standard gauge choice has order 2.

If ϕ_1 and ϕ_2 are both gauge choices of order 1, then their difference is of order 1 and has the form $D\psi$ for some 1-homogeneous ψ. Then ψ must be of order 0, and hence of the form $\psi = c(x)q_0$. It follows that an order-1 gauge choice is determined up to an additive term of the form $c'(x)q_0 + c(x)q_1$. More generally, by Section 7.2 two 1-homogeneous functions ϕ_1 and ϕ_2 of order 1 are equivalent if their difference has the linear form $a_0(x)q_0 + a_1(x)q_1$; and this is also necessary, since, again by Section 7.2, ϕ_2 must then have the form $\phi_1 + a(x)q_0 + c'(x)q_0 + c(x)q_1$.

8. Decomposition. The variational analysis has identified the form (39), where ϕ is 1-homogeneous, for a key local scoring rule. We now consider the properties of such a rule in more detail.

Starting from (39), we compute the expected score,

$$\begin{aligned} S(P,Q) &= \int_-^+ \mathrm{d}x\, p(x) S(x,Q) \\ &= \sum_{k\geq 0} (-1)^k \int_-^+ \mathrm{d}x\, p(x) \frac{\mathrm{d}^k}{\mathrm{d}x^k} \phi_{[k]}\{x, q(x), q'(x), \ldots, q^{(t)}(x)\} \end{aligned}$$

by evaluating the kth term in the sum using the integration by parts formula (16). Collecting terms, we obtain

$$S(P,Q) = S_0(P,Q) + S_+(P,Q) + S_-(P,Q), \tag{45}$$

where the *integral expected score* S_0 is given by

$$S_0(P,Q) = \int_-^+ \mathrm{d}x \sum_k p_k \phi_{[k]}(\mathbf{q}) \tag{46}$$

and

$$S_\pm(P,Q) = \mp S_b(\mathbf{p},\mathbf{q})|_\pm, \tag{47}$$

where the *boundary expected score* S_b is given by

$$S_b(\mathbf{p},\mathbf{q}) := \sum_{r\ge 0} p_r B_r \phi(\mathbf{q}) \tag{48}$$

with B_r defined by (32). In these formulas the dependence on x has been suppressed from the notation for simplicity, and we interpret $p_k := p^{(k)}(x)$, $q_k := q^{(k)}(x)$.

Correspondingly, the entropy $H(Q) = S(Q, Q)$ can be decomposed:

$$H(Q) = H_0(Q) + H_+(Q) + H_-(Q) \tag{49}$$

with *integral entropy*

$$H_0(Q) := \int_-^+ \mathrm{d}x \sum_k q_k \phi_{[k]}(\mathbf{q}) = \int_-^+ \mathrm{d}x\, \phi(\mathbf{q}), \tag{50}$$

where the last equality follows from Euler's theorem (38); and $H_\pm(Q) = \mp H_b(\mathbf{q})|_\pm$, where the *boundary entropy* $H_b(\mathbf{q})$ satisfies

$$\begin{aligned} H_b(\mathbf{q}) &= S_b(\mathbf{q},\mathbf{q}) \\ &= C\phi(\mathbf{q}) \end{aligned}$$

with the operator C defined by (33).

The divergence now becomes

$$d(P, Q) = d_0(P, Q) + d_+(P, Q) + d_-(P, Q), \tag{51}$$

where $d_0(P, Q) = S_0(P, Q) - H_0(P)$, etc. In particular, the boundary terms arise from the *boundary divergence*

$$d_b(P, Q) = \sum_r p_r B_r\{\phi(\mathbf{q}) - \phi(\mathbf{p})\} \tag{52}$$

[where the final term involves substituting $\mathbf{p}$ for $\mathbf{q}$ after computing $B_r\phi(\mathbf{q})$]; while, using (38), the *integral divergence* can be written as

$$d_0(P, Q) = \int_-^+ \mathrm{d}x \left[\left\{\phi(\mathbf{q}) + \sum_k (p_k - q_k)\phi_{[k]}(\mathbf{q})\right\} - \phi(\mathbf{p})\right]. \tag{53}$$

It is easily seen that both d_0 and d_b are unchanged by an equivalence transformation $\phi^* = \phi + \sum_{k\ge 0} a_k(x) q_k$.

8.1. *Change of gauge.* Although a key local scoring rule S is unchanged by a gauge transformation, the decompositions (45), (49) and (51), and in particular the expression (53) for d_0, typically do change, terms being redistributed between their constituents. Indeed, if we replace the generating ϕ by an alternative gauge choice

$$\phi^* = \phi + D\psi, \tag{54}$$

applying (46) yields

$$S_0^*(P, Q) = S_0(P, Q) + J$$

with

$$J := \int_-^+ \mathrm{d}x \sum_k p_k \left\{ \frac{\partial}{\partial q_k} D\psi(\mathbf{q}) \right\}.$$

Using $(\partial/\partial q_k)D = D(\partial/\partial q_k) + \partial/\partial q_{k-1}$, and the interpretation of D as $\mathrm{d}/\mathrm{d}x$, this reduces to

$$\begin{aligned} J &= \int_-^+ \mathrm{d}x \frac{\mathrm{d}}{\mathrm{d}x} \sum_k p_k \psi_{[k]}(\mathbf{q}) \\ &= \widehat{S}_+ + \widehat{S}_-, \end{aligned}$$

where $\widehat{S}_+ := \widehat{S}(\mathbf{p}, \mathbf{q})|_+$, $\widehat{S}_- := -\widehat{S}(\mathbf{p}, \mathbf{q})|_-$, with $\widehat{S}(\mathbf{p}, \mathbf{q}) := \sum_k p_k \psi_{[k]}(\mathbf{q})$.

Similarly, from (47) and (48) we find the boundary expected score transforming as

$$\begin{aligned} S_b^*(\mathbf{p}, \mathbf{q}) &= S_b(\mathbf{p}, \mathbf{q}) + \sum_k p_k B_k D\psi \\ &= S_b(\mathbf{p}, \mathbf{q}) + \widehat{S}(\mathbf{p}, \mathbf{q}) \end{aligned} \tag{55}$$

on using (35). The changes to the boundary terms thus compensate exactly (as they must) for the changes to the integral term.

We now have

$$H_0^*(P) = H_0(P) + \widehat{H}_+ + \widehat{H}_-$$

with $\widehat{H}_\pm := \pm \widehat{H}|_\pm$ and $\widehat{H}(\mathbf{p}) = \psi(\mathbf{p})$; this follows from (36) since ψ is 1-homogeneous. Correspondingly the boundary entropy transforms as $H_b^*(\mathbf{p}) = H_b + \psi(\mathbf{p})$.

It is notable that there is always a gauge choice for which the boundary entropy vanishes. Specifically:

THEOREM 8.1. *Let s be a key local score function. Then for the standard gauge choice $\phi = q_0 s$, the boundary entropy function H_b is identically* 0.

PROOF. From (37), $DH_b = DC\phi = E\phi - q_0\Lambda\phi$. Since ϕ is 1-homogeneous and $s = \Lambda\phi$, this becomes $\phi - q_0 s = 0$. So $0 = CDH_b = EH_b$ by (36). But $EH_b = H_b$ since H_b is 1-homogeneous. □

The effect on a gauge transformation on the decomposition of the divergence is

$$d_0^*(P, Q) = d_0(P, Q) + \widehat{d}_+ + \widehat{d}_-, \tag{56}$$

where $\widehat{d}_\pm := \pm\widehat{d}|_\pm$ with

$$\begin{aligned}\widehat{d}(\mathbf{p},\mathbf{q}) &= \sum_k p_k\psi_{[k]}(\mathbf{q}) - \psi(\mathbf{p}) \\ &= \psi(\mathbf{q}) + \sum_k (p_k - q_k)\psi_{[k]}(\mathbf{q}) - \psi(\mathbf{p})\end{aligned} \tag{57}$$

and with a compensating change to the boundary divergence d_b.

9. Propriety. In this section we investigate the propriety of a key local scoring rule S. The scoring rule S will be proper if and only if $d(P, Q) \geq 0$ for all P, $Q \in \mathcal{P}$. Clearly it is sufficient to require nonnegativity of each term in the right-hand side of the decomposition (51), and we proceed on this basis. We investigate d_+ and d_- in Section 10 below; here we consider the integral term d_0.

We note the similarity between formula (53) and that for the Bregman divergence (4) (especially where that is extended, as in Section 2.2, to allow further dependence of ϕ on x). Correspondingly, concavity of the defining function plays a crucial role here, too.

DEFINITION 9.1. We call a 1-homogeneous q-function $\phi(x,\mathbf{q})$ *concave* if, for every $x \in \mathcal{X}$, $\mathbf{q}_1, \mathbf{q}_2 \in \mathbf{Q}$,

$$\phi(x, \mathbf{q}_1 + \mathbf{q}_2) \leq \phi(x, \mathbf{q}_1) + \phi(x, \mathbf{q}_2) \tag{58}$$

(this is readily seen to be equivalent to the usual definition of concavity in $\mathbf{q}$, for each x); and *strictly concave* if strict inequality in (58) holds whenever the vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ are linearly independent.

THEOREM 9.1. *Suppose that the scoring rule S is generated by a concave 1-homogeneous q-function ϕ. Then $d_0(P, Q)$, as given by (53), is nonnegative. Further, if ϕ is strictly concave, then $d_0(P, Q) = 0$ if and only if $Q = P$.*

PROOF. Concavity implies that the integrand of (53) is nonnegative for each x; under strict concavity it will be strictly positive with positive probability when $Q \neq P$. $\square$

COROLLARY 9.2. *Suppose the conditions of Theorem 9.1 apply, and the boundary terms $d_+(P, Q)$ and $d_-(P, Q)$ in (51) vanish identically for $P, Q \in \mathcal{P}$. Then the (local, homogeneous) scoring rule (39) is proper (strictly proper if ϕ is strictly concave).*

9.1. *Checking concavity.* Given a 1-homogeneous q-function ϕ of order m, define, for $\mathbf{u} = (u_1, \dots, u_m) \in \mathbb{R}^m$,

$$\Phi(x, \mathbf{u}) := \phi(x, 1, \mathbf{u}). \tag{59}$$

Then $\phi(x, \mathbf{q})$ is determined by Φ:

$$\phi(x, \mathbf{q}) = q_0 \Phi(x, \mathbf{u}) \tag{60}$$

with $u_i = q_i/q_0$ $(i \geq 1)$. It is often easier to check concavity for Φ than for ϕ, and this is enough:

LEMMA 9.3. *Φ is concave in $\mathbf{u}$ if and only if ϕ is concave in $\mathbf{q}$.*

PROOF. "If" follows immediately from (59). Conversely, if Φ is concave,

$$\begin{aligned}\phi(x, \mathbf{p} + \mathbf{q}) &= (p_0 + q_0)\Phi\left(x, \frac{p_0}{p_0 + q_0}\frac{\mathbf{p}}{p_0} + \frac{q_0}{p_0 + q_0}\frac{\mathbf{q}}{q_0}\right) \\ &\geq p_0 \Phi\left(x, \frac{\mathbf{p}}{p_0}\right) + q_0 \Phi\left(x, \frac{\mathbf{q}}{q_0}\right) \\ &= \phi(x, \mathbf{p}) + \phi(x, \mathbf{q}). \qquad \square\end{aligned}$$

It is further easy to see that Φ is strictly concave in $\mathbf{u}$, in the usual sense, if and only if ϕ is strictly concave in $\mathbf{q}$ in the sense of Definition 9.1.

9.2. *Change of gauge.* Even if the initial gauge choice ϕ is concave in $\mathbf{q}$, so that $d_0(\mathbf{p}, \mathbf{q}) \geq 0$, under a gauge transformation (54) the term $\widehat{d}(\mathbf{p}, \mathbf{q})$, as given by (57), means that the gauge-transformed integral divergence term d_0^*, given by (56), need not be nonnegative; this would hold if the resulting gauge choice ϕ^* were itself concave, but typically this will not be so.

Note that if ψ in (54) is concave, then $\widehat{d}(\mathbf{p}, \mathbf{q}) \geq 0$. However, this does not ensure positivity of both additional terms, since while the added term $\widehat{d}_+ = \widehat{d}(\mathbf{p}, \mathbf{q})|_+$ will then be nonnegative, the other added term $\widehat{d}_- = -\widehat{d}(\mathbf{p}, \mathbf{q})|_-$ will be nonpositive.

EXAMPLE 9.1. The Hyvärinen scoring rule (9) on $\mathcal{X} = \mathbb{R}$ is generated by the strictly concave q-function $\phi = -\frac{1}{2}q_1^2/q_0$. Using this gauge choice in (53) yields [cf. (1)]

$$d_0(P, Q) = \frac{1}{2}\int \mathrm{d}x \, p(x)(v_1 - u_1)^2 \tag{61}$$

with $u_i := q^{(i)}(x)/q(x)$, $v_i := p^{(i)}(x)/p(x)$.

Alternatively we might use the standard gauge choice, $q_2 - \frac{1}{2}q_1^2/q_0$, which is also strictly concave, and indeed yields the same expression (61).

Now let $\psi := -\frac{1}{2} q_1 \ln(q_1/q_0)$, so that $D\psi = -\frac{1}{2}\{q_2 \ln(q_1/q_0) + q_2 - q_1^2/q_0\}$. Then $\phi^* = \phi + D\psi = -\frac{1}{2} q_2\{1 + \ln(q_1/q_0)\}$ is another possible gauge choice, generating the identical scoring rule S. However, ϕ^* is not concave, and the integral divergence term (53) associated with ϕ^* is

$$d_0^*(P, Q) = \frac{1}{2} \int \mathrm{d}x\, p(x) \left\{ u_2 \left(1 - \frac{v_1}{u_1} \right) + v_2 \ln \frac{v_1}{u_1} \right\},$$

which is not nonnegative. In this case the extra terms in (56) arise from $\widehat{d}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} p_0\{u_1 - v_1 + v_1 \ln(v_1/u_1)\}$.

In the light of the above example it might be conjectured that, if s can be generated from *some* concave gauge choice, then the standard gauge choice $\phi = q_0 s$ will be concave—equivalently, from Lemma 9.3, s itself will be a concave function of the $u_i = q_i/q_0$ $(i \geq 1)$—but this need not hold:

EXAMPLE 9.2. Take $\Phi = -u_1^4$ in (60). Then Φ, and hence ϕ, is concave, but $s = 12u_1^2 u_2 - 9u_1^4$ is not concave.

10. Boundary issues. The boundary divergence terms in (51) are $d_\pm(P, Q) = \mp d_b(\mathbf{p}, \mathbf{q})|_\pm$, where d_b is given by (52). Their behavior will depend on the family $\mathcal{P}$ of distributions under consideration, and specifically on the behavior, at the end-points $+$ and $-$, of the densities of distributions in $\mathcal{P}$.

For propriety of these terms, we want $d_b(\mathbf{p}, \mathbf{q})$ to be positive at the lower end-point $-$, and negative at the upper end-point $+$, for all $P, Q \in \mathcal{P}$. For simplicity we might impose conditions on $\mathcal{P}$ sufficient to ensure that, for all densities $p(\cdot)$, $q(\cdot) \in \mathcal{P}$, $d_b(\mathbf{p}, \mathbf{q})$ vanishes at the end-points. A family $\mathcal{P}$ having this property may be termed *valid* (with respect to the generating function ϕ). However, there does not appear to be a natural choice for such a valid class $\mathcal{P}$. In particular, if $\mathcal{P}$ and $\mathcal{P}'$ are both valid families, it does not follow that their union will be.

Note that the validity requirement depends on the gauge choice ϕ, and a change of gauge could assist in ensuring that it holds.

For the special case of the standard gauge choice, $\phi^* = q_0 s$, we know from Theorem 8.1 that the boundary entropy H_b^* vanishes. *If* the boundary quantities S_b^*, H_b^*, d_b^* behaved like regular quantities S, H, d we could deduce $d_b^* = 0$ [Dawid (1998)]; but this is a big "if," and the result will not hold without imposing further conditions.

10.1. *Second-order rules.* For a second-order rule with 1-homogeneous generator $\phi(x, q_0, q_1)$, we find

$$d_b = p_0\{\phi_{[1]}(\mathbf{q}) - \phi_{[1]}(\mathbf{p})\}.$$

Alternatively, the standard gauge choice is $\phi^* = \phi + D\psi$ with $\psi = -C\phi = -q_0\phi_{[1]}$. From Section 8.1, we find

$$d_b^*(\mathbf{p}, \mathbf{q}) = S_b^*(\mathbf{p}, \mathbf{q}) = -q_0\big(p_0\phi_{[01]} + p_1\phi_{[11]}\big).$$

That this vanishes (as we know from Theorem 8.1 it must) for $\mathbf{p} = \mathbf{q}$ may be seen on differentiating the relation $\phi = q_0\phi_{[0]} + q_1\phi_{[1]}$ with respect to q_1; that it does not depend on the choice of gauge ϕ of order 1 follows from Section 7.4.

With $p_i = p^{(i)}(x)$, etc., we want $d_b(\mathbf{p}, \mathbf{q})$ [or, for the standard gauge choice, $d_b^*(\mathbf{p}, \mathbf{q})$] to vanish in the limit as we approach the end-points $-$ and $+$, for all densities $p(\cdot)$, $q(\cdot)$ of distributions in $\mathcal{P}$. Conditions for validity will thus involve the behavior of $p(x)$ and $p'(x)$ at these end-points.

For example, for the Hyvärinen rule, with gauge choice $\phi = -\frac{1}{2}q_1^2/q_0$, we require

$$d_b(\mathbf{p}, \mathbf{q}) = p_0\left(\frac{p_1}{p_0} - \frac{q_1}{q_0}\right) \to 0 \tag{62}$$

as we approach the end-points of $\mathcal{X}$. (The same expression for d_b arises if we use the standard gauge choice $\phi^* = q_2 - \frac{1}{2}q_1^2/q_0$, which in this case is equivalent to ϕ.) To ensure (62) we might require, for example, that, for all densities $p(\cdot)$ in $\mathcal{P}$, $\lim_{x\to\pm} p(x) = 0$ and $\lim_{x\to\pm} p'(x)/p(x)$ is finite. However, this excludes the possibility that both p and q are normal densities on $\mathcal{X} = \mathbb{R}$, even though, with this choice, d_b as given by (62) does vanish at $\pm\infty$. Ehm and Gneiting (2010, 2012) described alternative conditions on $\mathcal{P}$ that do admit this case.

In the ideal situation we will have a (strictly) concave 1-homogeneous ϕ, and a family $\mathcal{P}$ valid with respect to ϕ. Then the associated key local scoring rule S will be (strictly) proper.

11. Transformation of the data. So far we have considered a variable X taking values in a real interval $\mathcal{X}$, and have made essential use of the Euclidean structure of $\mathcal{X}$ to define probability densities, derivatives, etc. Taking a step backward, suppose we start with an abstract measurable sample space (the *basic sample space*) $\mathcal{X}^*$, a *basic variable* X^* taking values in $\mathcal{X}^*$, and a collection $\mathcal{P}^*$ of *basic distributions* for X^* over $\mathcal{X}^*$. Without assuming any further structure, we can define a *basic scoring rule* $S^*: \mathcal{X}^* \times \mathcal{P}^* \to \mathbb{R}$, and introduce the property of (strict) propriety, exactly as before. However, at this level of generality it is less straightforward to define what we should mean by saying that a basic scoring rule is *local*. To do this we proceed as follows.

We suppose given a collection $\Xi = \{\xi\}$ of *charts*, where each ξ is an invertible measurable function from $\mathcal{X}^*$ onto some open interval $\mathcal{X} \subseteq \mathbb{R}$, and such that, for $\xi, \overline{\xi} \in \Xi$, the composition $\overline{\xi}\xi^{-1}: \mathcal{X} \to \overline{\mathcal{X}}$ is *smooth* and *regular*, that is, infinitely often differentiable with strictly positive first derivative. In other words, the basic space is a one-dimensional simply connected smooth manifold.

Picking any specific chart ξ produces a concrete *representation* of the abstract basic structure, in terms of the real variable $X := \xi(X^*)$, and, for any $Q^* \in \mathcal{P}^*$, the induced distribution Q for X on $\mathcal{X} \subseteq \mathbb{R}$ [so that $Q(A) = Q^*\{\xi^{-1}(A)\}$]; we take $\mathcal{P} := \{P : P^* \in \mathcal{P}^*\}$. Correspondingly, a *basic function* $f^* : \mathcal{X}^* \times \mathcal{P}^* \to \mathbb{R}$ (e.g., a scoring rule) is *represented* by $f : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$, such that $f(x, Q) = f^*(x^*, Q^*)$.

Let $\xi, \overline{\xi}$ be two such charts, and $X = \xi(X^*)$, $\overline{X} = \overline{\xi}(X^*)$, etc. Then $\overline{X} = \gamma(X)$, where $\gamma = \overline{\xi}\xi^{-1}$ is strictly increasing, and both γ and $\delta := \gamma^{-1}$ are smooth and regular. A given basic distribution Q^* for X^* can be represented *either* by the distribution Q, for X, *or* by $\overline{Q}$, for $\overline{X}$. We assume that Q has a density function, $q(\cdot)$, with respect to Lebesgue measure on $\mathcal{X}$; then the density function $\overline{q}(\cdot)$ of $\overline{Q}$ with respect to Lebesgue measure on $\overline{\mathcal{X}}$ will likewise exist, and, with $\overline{x} = \gamma(x)$, we will have

$$\overline{q}(\overline{x}) = q(x)\,\frac{\mathrm{d}x}{\mathrm{d}\overline{x}} = \alpha(x) q(x) \tag{63}$$

with $\alpha(x) := \gamma'(x)^{-1}$. An easy induction shows that we can express

$$\overline{q}^{(k)}(\overline{x}) = \overline{T}_k\bigl(x, q(x), \ldots, q^{(k)}(x)\bigr), \tag{64}$$

where $\overline{T}_k$ has the form

$$\overline{T}_k(x, q_0, \ldots, q_k) = \sum_r a_{kr}(x) q_r \tag{65}$$

and the coefficients $a_{kr}(x)$ satisfy $a_{kr}(x) = 0$ unless $0 \le r \le k$, $a_{00}(x) = \alpha(x)$, and

$$a_{k+1,r}(x) = \alpha(x)\{a'_{kr}(x) + a_{k,r-1}(x)\}. \tag{66}$$

In similar fashion we can express

$$\begin{aligned} q^{(k)}(x) &= T_k\bigl(\overline{x}, \overline{q}(\overline{x}), \ldots, \overline{q}^{(k)}(\overline{x})\bigr) \\ &= \sum_r \overline{a}_{kr}(\overline{x}) \overline{q}^{(r)}(\overline{x}). \end{aligned} \tag{67}$$

It readily follows from (64) and (67) that a basic function $f^*(x^*, Q^*)$ can be written, in the ξ-representation, in the form $f(x, q(x), q'(x), \ldots, q^{(m)}(x))$ if and only if the analogous property holds in the $\overline{\xi}$-representation: $f^*(x^*, Q^*) = \overline{f}(\overline{x}, \overline{q}(\overline{x}), \overline{q}'(\overline{x}), \ldots, \overline{q}^{(m)}(\overline{x}))$. That is, the property of being *m-local* is independent of the particular representation used. When this property holds for one, and thus for all, representations, we can say that the basic function $f^*(x^*, Q^*)$ itself is *m-local*; a q-function f such that $f^*(x^*, Q^*) = f(x, q(x), q'(x), \ldots, q^{(m)}(x))$ is the *ξ-representation* of f^*. We denote the vector space of all local basic functions by $\mathcal{V}^*$.

At a more abstract level, motivated by (64) and (65), we define variables

$$\begin{aligned} \overline{x} &:= \gamma(x), \\ \overline{q}_k &:= \overline{T}_k(x, q_0, \ldots, q_k) = \sum_r a_{kr}(x) q_r. \end{aligned} \tag{68}$$

Inversely, we will then have

$$x = \delta(\overline{x}),$$
$$q_k = T_k(\overline{x}, \overline{q}_0, \ldots, \overline{q}_k) = \sum_r \overline{a}_{kr}(\overline{x})\overline{q}_r. \tag{69}$$

Using (69), any q-function of order m, $f(x, q_0, \ldots, q_m)$ can be rewritten as $\overline{f}(\overline{x}, \overline{q}_0, \ldots, \overline{q}_m)$. If $f^* \in \mathcal{V}^*$ has ξ- and $\overline{\xi}$-representations f and $\overline{f}$, respectively, then $\overline{f}$ can be obtained by reexpressing f in this way. Since T_k is 1-homogeneous, f is homogeneous of degree h in the q's if and only if $\overline{f}$ is homogeneous of degree h in the $\overline{q}$'s. In this case we may term the underlying local basic function $f^* \in \mathcal{V}^*$ *h-homogeneous*. Likewise, since (for fixed x or $\overline{x}$) the functions T_k and $\overline{T}_k$ are linear, f is (strictly) concave in the q's if and only if $\overline{f}$ is (strictly) concave in the $\overline{q}$'s—in which case we may term f^* itself (*strictly*) *concave*.

11.1. *Invariant operators*. The linear differential operators D and L have only been defined in terms of a specific representation of the problem on the real line, as determined by some chart ξ. Applying these definitions starting from a different real representation, determined by a chart $\overline{\xi}$, we will obtain possibly different operators, $\overline{D}$, $\overline{L}$. The following results relate these. We need the following lemma:

LEMMA 11.1. *We have*

$$\frac{\partial}{\partial q_r} = \sum_k a_{kr} \frac{\partial}{\partial \overline{q}_k}, \tag{70}$$

$$\frac{\partial}{\partial x} = \alpha^{-1} \frac{\partial}{\partial \overline{x}} + \sum_r q_r \sum_k a'_{kr} \frac{\partial}{\partial \overline{q}_k}. \tag{71}$$

PROOF. Equation (70) follows immediately from (68). For (71) we have

$$\frac{\partial}{\partial x} = \frac{\mathrm{d}\overline{x}}{\mathrm{d}x} \frac{\partial}{\partial \overline{x}} + \sum_k \frac{\partial \overline{q}_k}{\partial x} \frac{\partial}{\partial \overline{q}_k}.$$

But $\mathrm{d}\overline{x}/\mathrm{d}x = \alpha^{-1}$, while from (68)

$$\frac{\partial \overline{q}_k}{\partial x} = \sum_r a'_{kr} q_r,$$

so (71) follows. □

We now show that if f and $\overline{f}$ are, respectively, the ξ and $\overline{\xi}$ representations of the same basic function f^*, then $\overline{Df}$ is the $\overline{\xi}$-representation of the basic function whose ξ-representation is $\alpha(x)Df$. Note that the function α, and hence the basic function so represented, will depend on the charts considered.

THEOREM 11.2. *It holds that*

$$\overline{D} = \alpha(x) D.$$

PROOF. Informally, we observe that D corresponds to the total derivative $\mathrm{d}/\mathrm{d}x$ and $\overline{D}$ to $\mathrm{d}/\mathrm{d}\overline{x}$. Thus we expect $\overline{D} = (\mathrm{d}x/\mathrm{d}\overline{x}) D$.

More formally, we have

$$\begin{aligned}\overline{D} &= \frac{\partial}{\partial \overline{x}} + \sum_k \overline{q}_{k+1} \frac{\partial}{\partial \overline{q}_k} \\ &= \frac{\partial}{\partial \overline{x}} + \sum_r q_r \sum_k a_{k+1,r} \frac{\partial}{\partial \overline{q}_k}\end{aligned}$$

on using (68). From (66) this is

$$\frac{\partial}{\partial \overline{x}} + \alpha \sum_r q_r \sum_k (a'_{k,r} + a_{k,r-1}) \frac{\partial}{\partial \overline{q}_k}.$$

On applying Lemma 11.1 this reduces to αD. $\square$

Since, by the transformation rule (63) for densities, $\overline{q}_0 = \alpha(x) q_0$, we thus have

COROLLARY 11.3. *It holds that* $\overline{q}_0^{-1}\overline{D} = q_0^{-1} D$.

It follows from Corollary 11.3 that, for $f^* \in \mathcal{V}^*$, there exists $g^* \in \mathcal{V}^*$ such that, in any representation, $g = q_0^{-1} Df$. This shows the existence of an "invariant" linear operator D^* on $\mathcal{V}^*$ such that, in any representation, $D^* f^*$ is represented by $q_0^{-1} Df$.

We next show that there exists an invariant linear operator L^* on $\mathcal{V}^*$ such that, in any representation, if f^* is represented by f, then $L^* f^*$ is represented by Lf.

THEOREM 11.4. *We have* $\overline{L} = L$.

PROOF. On substituting (63) and (70) into the definition (19) of L and rearranging, we obtain

$$L = \sum_k (-1)^{k+1} A_k \alpha^{-1} \overline{q}_0 \frac{\partial}{\partial \overline{q}_k}, \tag{72}$$

where the operator A_k is given by

$$A_k = \sum_r (-1)^{k-r} D^r a_{kr} \circ.$$

The theorem will thus be proved if we can show $A_k = \overline{D}^k \alpha \circ$; that is, using Theorem 11.2, we have to show:

$$H_k : (\alpha D)^k \alpha \circ = \sum_r (-1)^{k-r} D^r a_{kr} \circ. \tag{73}$$

We prove (73) by induction on k. First, H_0 holds since both sides reduce to $\alpha\circ$. Now suppose H_k holds. Then

$$(\alpha D)^{k+1}\alpha\circ = (\alpha D)^k \alpha D\alpha\circ$$
$$= \sum_r (-1)^{k-r} D^r a_{kr} D\alpha\circ. \tag{74}$$

But $a_{kr}D = (Da_{kr}\circ) - a'_{kr}\circ$, so that (74) becomes

$$\sum_r (-1)^{k-r} D^{r+1} a_{kr}\alpha\circ - \sum_r (-1)^{k-r} D^r a'_{kr}\alpha\circ,$$

which can be written as

$$\sum_r (-1)^{k+1-r} D^r \alpha(a'_{kr} + a_{k,r-1})\circ$$

and on applying (66) we have verified H_{k+1}. □

11.2. *Invariance of scoring rule.* On applying Theorem 11.4, we see that the general homogeneous key local scoring rule, as given by (ii) of Theorem 6.4, can be expressed invariantly as

$$S^*(x^*, Q^*) = (I - L^*)g^*(x^*, Q^*),$$

where g^* is a 0-homogeneous local basic function. Then, in any representation, we will have $S(x, Q) = (I - L)g(x, Q)$. We may thus say that the scoring rule S^* is *derived from* the local basic function g^*. In particular the expected score $S^*(P^*, Q^*)$, and consequently the entropy function $H^*(P^*)$ and the divergence function $d^*(P^*, Q^*)$, are fully determined by the basic function g^*, independently of how that may be represented.

In fact more is true: the individual components $S_0(P, Q)$, $S_+(P, Q)$, $S_-(P, Q)$ of $S(P, Q)$, in the decomposition (45) arising from the integration by parts, each correspond to an invariant expression $S_0^*(P^*, Q^*)$, $S_+^*(P^*, Q^*)$, $S_-^*(P^*, Q^*)$ (and similarly for the decompositions of H and d).

We show this first for the integral term S_0. We need the following lemma, showing that the expression $\Pi := \sum_r p_r\, \partial/\partial q_r$ represents an invariant operator Π^* (depending on a distribution P^*, and acting on a function of a distribution Q^*, both defined over $\mathcal{V}^*$).

LEMMA 11.5. *We have*

$$\sum_r p_r \frac{\partial}{\partial q_r} = \sum_k \overline{p}_k \frac{\partial}{\partial \overline{q}_k}.$$

PROOF. Using (70) and (68), we have

$$\sum_r p_r \frac{\partial}{\partial q_r} = \sum_k \frac{\partial}{\partial \overline{q}_k} \sum_{r=0}^{k} a_{kr}(x) p_r$$
$$= \sum_k \overline{p}_k \frac{\partial}{\partial \overline{q}_k}. \qquad \square$$

Now consider expression (46) for $S_0(P, Q)$, where, in accordance with (ii) and (iii) of Theorem 6.4, $\phi = q_0 g$, with g the representation of a local basic function g^*. The integrand can then be written as $(p_0 + q_0 \Pi)g$, whence $S_0(P, Q) = \mathrm{E}_P g + \mathrm{E}_Q(\Pi g) = \mathrm{E}_{P^*} g^* + \mathrm{E}_{Q^*}(\Pi^* g^*)$—which thus has an invariant form, $S_0^*(P^*, Q^*)$, independently of the representation employed.

We next demonstrate the corresponding property for the boundary term S_b.

THEOREM 11.6. *We have*

$$\sum_r p_r B_r = \sum_k \overline{p}_k \overline{B}_k \alpha \circ. \tag{75}$$

PROOF. On substituting (70), using (65) and Theorem 11.2, and rearranging, the statement of the theorem becomes

$$\sum_r p_r \sum_{m \geq r+1} \sum_{k=r+1}^{m} (-1)^{k-1-r} D^{k-1-r} a_{mk} \frac{\partial}{\partial \overline{q}_m}$$
$$= \sum_r p_r \sum_{m \geq r+1} \sum_{k=r}^{m-1} a_{kr} (-1)^{m-1-k} (\alpha D)^{m-1-k} \frac{\partial}{\partial \overline{q}_m} \alpha \circ.$$

The theorem will thus be proved if we can show

$$H_m^{(r)}: \sum_{k=r+1}^{m} (-1)^{k-1-r} D^{k-1-r} a_{mk} \circ = \sum_{k=r}^{m-1} a_{kr} (-1)^{m-1-k} (\alpha D)^{m-1-k} \alpha \circ \tag{76}$$

for all $m \geq r+1$. We prove (76) by induction on m. First, $H_{r+1}^{(r)}$ holds since the left-hand side reduces to $a_{r+1,r+1} \circ$ and the right-hand side reduces to $a_{rr} \alpha \circ$, and these are equal by (66). Now suppose $H_m^{(r)}$ holds. Then

$$\sum_{k=r+1}^{m} (-1)^{k-1-r} D^{k-1-r} a_{mk} D \alpha \circ \tag{77}$$

$$= \sum_{k=r}^{m-1} a_{kr} (-1)^{m-1-k} (\alpha D)^{m-1-k} \alpha D \alpha \circ. \tag{78}$$

But $a_{mk}D = (Da_{mk}\circ) - a'_{mk}\circ$, so that, on applying (66), the left-hand side of (78) becomes

$$a_{mr}\alpha\circ - \sum_{k=r+1}^{m+1} (-1)^{k-1-r} D^{k-1-r} a_{m+1,k}\circ.$$

The right-hand side of (78) straightforwardly becomes

$$a_{mr}\alpha\circ - \sum_{k=r}^{m} a_{kr}(-1)^{m-k}(\alpha D)^{m-k}\alpha\circ,$$

and we have thus verified H_{m+1}. $\square$

It follows from (75) that $\sum_r p_r B_r q_0 = \sum_r \overline{p}_r \overline{B}_r \overline{q}_0$, which thus defines an invariant operator. Let now S^*, with representations S, $\overline{S}$, derive from the 0-homogeneous basic function g^*, with representations g, $\overline{g}$. On using (48), in which $\phi = q_0 g$, we get

$$S_b(\mathbf{p}, \mathbf{q}) = \sum_r p_r B_r q_0 g = \sum_r \overline{p}_r \overline{B}_r \overline{q}_0 \overline{g} = \overline{S}_b(\overline{\mathbf{p}}, \overline{\mathbf{q}}).$$

Hence by (47) the boundary contributions $S_{\pm}(P, Q)$ will be the same in all representations.

EXAMPLE 11.1 (Modified Hyvärinen rule). Take $\mathcal{X} = (0, \infty)$, $\overline{\mathcal{X}} = \mathbb{R}$, $\gamma(x) \equiv \ln x$ (so $\overline{X} = \ln X$). Then $\alpha(x) \equiv x$ and we find $\overline{q}_0 = xq_0$, $\overline{q}_1 = xq_0 + x^2 q_1$, $\overline{q}_2 = xq_0 + 3x^2 q_1 + x^3 q_2$.

Let the scoring rule in the $\overline{\xi}$-representation, $\overline{S}$, be defined by the Hyvärinen formula:

$$\overline{S}(\overline{x}, \overline{Q}) = \frac{\overline{q}''(\overline{x})}{\overline{q}(\overline{x})} - \frac{1}{2}\left\{\frac{\overline{q}'(\overline{x})}{\overline{q}(\overline{x})}\right\}^2. \tag{79}$$

This derives from the function $\overline{g} = -\frac{1}{2}(\overline{q}_1/\overline{q}_0)^2$.

Reexpressed in the ξ-representation, we have

$$S(x, Q) = x^2\left[\frac{q''(x)}{q(x)} - \frac{1}{2}\left\{\frac{q'(x)}{q(x)}\right\}^2\right] + 2x\frac{q'(x)}{q(x)} + \frac{1}{2}, \tag{80}$$

which itself derives from the ξ-reexpression of $\overline{g}$, viz., $g = -\frac{1}{2}(1 + xq_1/q_0)^2$. That is, it is generated by $\phi = q_0 g = -\frac{1}{2}q_0 - xq_1 - \frac{1}{2}x^2 q_1^2/q_0$. The simpler choice $\phi^* = -\frac{1}{2}x^2 q_1^2/q_0$ is equivalent to ϕ, and thus generates an equivalent scoring rule, with the same divergence function; in fact, it simply eliminates the final term $+\frac{1}{2}$ in (80). This form of the scoring rule also appears in equation (28) of Hyvärinen (2007).

For this scoring rule, a class $\mathcal{P}^*$ of distributions for the basic variable X^* will be valid if, for $P, Q \in \mathcal{P}^*$, $\overline{p}_0\{(\overline{p}_1/\overline{p}_0) - (\overline{q}_1/\overline{q}_0)\} \to 0$ as $\overline{x} \to \pm\infty$, where these expressions are based on the $\overline{\xi}$-representation [in which $\overline{S}$ is given by (79)]. Reexpressing this in the ξ-representation, we want

$$x^2 p_0\{(p_1/p_0) - (q_1/q_0)\} \to 0 \qquad \text{as } x \to 0 \text{ or } \infty. \tag{81}$$

At the lower end-point 0 of $\mathcal{X}$, this condition is less restrictive than the corresponding condition (62) for the regular Hyvärinen scoring rule defined directly on $\mathcal{X}$—although it becomes more restrictive at ∞.

In particular, suppose we consider the family $\mathcal{E}$ of exponential densities:

$$q(x|\theta) = \theta e^{-\theta x} \qquad (x, \theta > 0).$$

For $p, q \in \mathcal{E}$, condition (81) is satisfied, whereas (62) is not. If we tried to apply the unmodified Hyvärinen score (9) to estimate θ in this model, we would obtain $S(x, Q_\theta) = \frac{1}{2}\theta^2$, and (7) would then appear to yield the clearly nonsensical estimate $\widehat{\theta} \equiv 0$. This is due to failure of the boundary conditions, so that the original Hyvärinen rule is not in fact proper in this case. The modified rule (80) is proper for this family, and yields the consistent estimator $2\sum_i X_i / \sum_i X_i^2$.

12. Discussion and further work. In this paper we have investigated local scoring rules only for the case that the sample space is an open interval on the real line. The general ideas extend to the case that the sample space is a simply-connected d-dimensional differentiable manifold. This raises challenging new technical problems, but could deliver a fundamentally improved understanding and illuminate issues associated with boundary problems.

Acknowledgments. We are grateful to Valentina Mameli for identifying errors and simplifying proofs in an earlier version of this work. In addition we have benefited from helpful discussions with Werner Ehm, Tilmann Gneiting and Peter Whittle.

REFERENCES

BARNDORFF-NIELSEN, O. (1976). Plausibility inference. *J. Roy. Statist. Soc. Ser. B* **38** 103–131.

BERNARDO, J.-M. (1979). Expected information as expected utility. *Ann. Statist.* **7** 686–690. MR0527503

BREGMAN, L. M. (1967). The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. and Math. Phys.* **7** 200–217.

CSISZÁR, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.* **19** 2032–2066. MR1135163

DAWID, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson and C. B. Read, eds.) **7** 210–218. Wiley, New York.

DAWID, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Dept. Statistical Science, Univ. College London. Available at http://www.ucl.ac.uk/statistics/research/pdfs/139b.zip.

DAWID, A. P. (2007). The geometry of proper scoring rules. *Ann. Inst. Statist. Math.* **59** 77–93. MR2396033

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* 22–28. Univ. Tokyo, Tokyo, Japan.

DAWID, A. P., LAURITZEN, S. and PARRY, M. (2012). Proper local scoring rules on discrete sample spaces. *Ann. Statist.* **40** 593–608.

EGUCHI, S. (2008). Information divergence geometry and the application to statistical machine learning. In *Information Theory and Statistical Learning* (F. Emmert-Streib and M. Dehmer, eds.) 309–332. Springer, New York.

EHM, W. and GNEITING, T. (2010). Local proper scoring rules. Technical Report 551, Dept. Statistics, Univ. Washington. Available at http://www.stat.washington.edu/research/reports/2009/tr551.pdf.

EHM, W. and GNEITING, T. (2012). Local proper scoring rules of order two. *Ann. Statist.* **40** 609–637.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York. MR0829458

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

HYVÄRINEN, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6** 695–709 (electronic). MR2249836

HYVÄRINEN, A. (2007). Some extensions of score matching. *Comput. Statist. Data Anal.* **51** 2499–2512. MR2338984

SAVAGE, L. J. (1954). *The Foundations of Statistics*. Wiley, New York. MR0063582

TROUTMAN, J. L. (1983). *Variational Calculus with Elementary Convexity*. Springer, New York. MR0697723

VAN BRUNT, B. (2004). *The Calculus of Variations*. Springer, New York. MR2004181

M. PARRY
DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF OTAGO
P.O. BOX 56
DUNEDIN 9054
NEW ZEALAND
E-MAIL: mparry@maths.otago.ac.nz

A. P. DAWID
STATISTICAL LABORATORY
CENTRE FOR MATHEMATICAL SCIENCES
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD, CAMBRIDGE CB3 0WB
UNITED KINGDOM
E-MAIL: A.P.Dawid@statslab.cam.ac.uk

S. LAURITZEN
DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
SOUTH PARKS ROAD
OXFORD OX1 3TG
UNITED KINGDOM
E-MAIL: steffen@stats.ox.ac.uk
URL: http://www.foo.com