Minimum proper loss estimators for parametric models

Matthew J. Holland* Kazushi Ikeda

Nara Institute of Science and Technology
Ikoma, Nara, Japan

Abstract

In this paper, we propose a methodology for systematically deriving estimators minimizing proper loss functions defined on parametric statistical models, by restricting ourselves to losses taking a functional form which allows for straightforward proofs of key properties. Proving propriety is tantamount to deriving a pairwise divergence quantity between probability measures, admitting a natural interpretation of minimum proper loss estimators as divergence minimizers. We show that proper losses of varying complexity can be readily constructed given propriety of losses taking a rudimentary form, and that for many important models, verifying desired properties in the rudimentary case is immediate. As a special case, we derive computationally tractable estimators requiring only the first and second moments, verify strict propriety properties, and empirically confirm their utility though parameter estimation tasks using both controlled simulations and real-world meteorological network data sets. Comparisons against numerous standard estimators show that the proposed estimators are at least competitive with, and often markedly superior to all standard references, uniformly across tasks, data sets, and model classes, suggesting a strong alternative to standard benchmarks in a wide variety of estimation problems.

1 Introduction

Estimation tasks in systems with stochastic elements are typically formulated as optimizations of a particular objective (henceforth, a "loss" to be minimized), namely a function defined over the collection of candidate probability distributions (a probabilistic "model") the system may select from. Distributions are often specified by a finite collection of parameters, and the "optimality" of any parameter estimate is always with respect to a given loss. Given a model optimizing the parameters for any particular loss is relatively straightforward; selecting an appropriate loss, however, is far from obvious. In this paper we seek a systematic means of deriving loss functions with desirable properties for large classes of important parametric models. We begin with a review of relevant literature and a description of the contributions of this paper given in context.

Let us start with literature on loss functions. Our interest is typically in an observable random vector $\boldsymbol{x}$ on underlying space $(\Omega, \mathcal{A}, \mathbf{P})$, and making inferential statements about its distribution $P_X(B) := \mathbf{P}\{\boldsymbol{x} \in B\}$ over the Borel sets of $\mathbb{R}^d$. Given model $\mathcal{P}$, a collection of probabilities, any reasonable loss will be a data-dependent evaluation of our estimation. One natural formulation is to define real-valued losses $\lambda(\cdot, Q)$ which are measurable for each $Q \in \mathcal{P}$, and given N independent copies $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ consider, for example, the minimization of

*Email: matthew-h@is.naist.jp.

random quantity $\sum_{n=1}^{N} \lambda(\boldsymbol{x}_n, Q)$ with respect to $Q \in \mathcal{P}$. The quality of such an estimate will depend on λ and $\mathcal{P}$, and much relevant theory seeks to verify useful attributes of important combinations of loss/model classes. "Propriety" (defined in Section 2) of loss functions says that the true distribution uniquely minimizes the expected loss, and has been the subject of much analysis in recent years [19, 16, 13, 35]. From comprehensive work by Gneiting and Raftery [19], when $\mathcal{P}$ is convex, under regularity conditions one may characterize $\mathcal{P}$-proper loss functions as those which can be expressed as $-\lambda(\boldsymbol{x}, P) = g(P) + (g^*(\boldsymbol{x}, P) - \mathbf{E}_P g^*(\boldsymbol{x}, P))$ where g has a convexity property and g^* is analogous to the sub-gradient of convex functions on $\mathbb{R}^d$. More general classes of losses are studied in the context of statistical decision problems in Dawid and Sebastiani [14], with a geometric look explicitly at proper losses by Dawid [12]. Recent work by Ehm and Gneiting [16] looks at "locality" in losses, where at each $\boldsymbol{x} \in \mathbb{R}^d$, $\lambda(\boldsymbol{x}, Q)$ depends on (when existent) density q of Q only in the form $q(\boldsymbol{x})$. The (negative) log-likelihood is known to be the only local proper loss, but in Ehm and Gneiting [16] and Dawid et al. [13] they weaken locality conditions, allowing dependence on the kth derivative $q^{(k)}$, and analyse classes of k-local proper loss functions.

Considering the referred works above, we offer sufficient conditions for $\mathcal{P}$-propriety in the case where $\mathcal{P}$ has a parametrization to some $\Theta \subset \mathbb{R}^d$, which maintain the (convex function) + (function of sub-gradient) notion from the general characterization due to Gneiting and Raftery [19], while being more readily verifiable for the model classes of interest. We propose a general functional form to which these conditions may be applied, giving a systematic approach to deriving proper loss functions for a given model class. In addition, for important special cases, propriety proofs for rudimentary losses implies propriety for compositions of these simple losses, allowing additional freedom in estimator construction.

Next we look at the closely related literature on estimation via minimum pairwise dissimilarity. If λ is a (strictly) proper loss, it furnishes a valid "divergence" or "contrast" metric (defined in Section 2) by $d(P, Q) \coloneqq \mathbf{E}_Q(\lambda(\cdot, P) - \lambda(\cdot, Q))$ on $\mathcal{P}$ which satisfies non-negativity and definiteness. This lets one consider the λ-optimal estimator as a divergence-minimizing estimator. As above, negative log-likelihood yields the KL divergence, and well-known applications include image segmentation and texture retrieval [36], blind source separation (BSS) [38], non-linear denoising [1], and adaptive control of continuous-state systems [31]. In Pfanzagl [37], conditions for the existence of minimum contrast estimators are given, and Birgé and Massart [8] and the references within modify consistency results for M-estimators, originally due to Huber [26], to the minimum contrast case. In Eguchi [15], using differential geometric properties of exponential families of densities, under smoothness and convexity constraints, estimators minimizing divergences of the form $\int q(\boldsymbol{x}) f(p(\boldsymbol{x})/q(\boldsymbol{x}))\tau(d\boldsymbol{x})$ are shown to converge efficiently in the Fisher information sense, noting p, q denotes densities of some $P, Q \in \mathcal{P}$. If τ is a dominating measure of all $Q \in \mathcal{P}$, we note that this essentially gives us the f-divergence $d_f(P, Q) \coloneqq \mathbf{E}_Q f(p/q)$ of Csiszár [10], whereany real f convex on $(0, \infty)$ and strictly convex at 1 is admissible. This class of divergences has many useful properties including convexity and invariance to bijective transformations [11], and continues to be the subject of analysis. Letting f^* be the convex conjugate of f, Broniatowski and Keziou [9] use a result showing that when f is differentiable, $d_f(P, Q)$ (under the name ϕ-divergence) can be dually expressed as the supremum of $\mathbf{E}_P g - \mathbf{E}_Q f^*(g)$ with g spanning a sufficiently large class of Borel measurable functions. They also look at asymptotic properties of minimum d_f estimators that hold uniformly over special function classes. In Basu et al. [6], consistency and efficiency results for estimators minimizing the "density power divergence" $d_\alpha(p, q)$ between two density functions, parametrized by $\alpha > 0$, are given. Note that given a sample from P with density p, minimizing $d_\alpha(p, q)$ over q is equivalent to minimizing $\int q^{1+\alpha}(x)\, dx - (1 + \alpha^{-1})\mathbf{E}_P q^\alpha$, and does not readily decompose into the expected difference of two proper losses.

As stated after the review of loss function literature, our contributions relate explicitly to principled proofs of loss function propriety. While every proper loss function generates a valid divergence, not all valid divergences need decompose into proper losses. As such, the analysis in this paper can be considered to equivalently describe a sub-class of divergences defined on models with an appropriate parametrization. Consistency arguments for minimum proper loss estimators can be made using the referred literature on minimum contrast estimators, so long as a given loss has been shown to be (strictly) proper. Verification of the required regularity conditions is a model-specific technical exercise that falls outside the scope of this paper. Broad efficiency results for curved exponential families are given by Eguchi [15], and while our results imply strict propriety of particular loss classes on such models, in general these do not result in special cases of the f-divergence, which is presupposed for many key results. We reiterate that the current literature as discussed above relates to analysis of functions already known to be proper losses (or similarly analysis of divergences generated from proper losses), while it is our aim to contribute to a principled derivation of valid proper loss functions given parametric models under weak conditions.

In Section 2 we begin with preliminary definitions and notation, and then elucidate the details of the proposed approach. Key results evaluating method performance in several tasks using real-world data and a discussion of results is given in Section 3. The paper concludes in Section 4, with closing remarks and a discussion of possible future lines of work.

2 Methods

We begin by establishing notation and giving key preliminary definitions. This is followed by the main results of this paper accompanied by motivating examples. The Section closes with a detailed application to a concrete parameter estimation task.

2.1 Preliminaries

Observations will be random vectors $\boldsymbol{x} := \boldsymbol{x}(\omega) \in \mathbb{R}^d$ on some underlying space $(\Omega, \mathcal{A}, \mathbf{P})$. Let $\mathcal{B}_d$ denote the Borel sets of $\mathcal{R}^d$. We are solely interested in inferring traits of $P_X(B) := \mathbf{P}\{\boldsymbol{x} \in B\}, B \in \mathcal{B}_d$. While this distribution is unknown, we do observe $\boldsymbol{x} \sim P_X$, and based on these observations we evaluate potential candidate distributions. A *model* then, here denoted $\mathcal{P}$, will refer to a class of valid probabilities on $\mathcal{B}_d$. To carry out this evaluation, we make use of $\lambda(\cdot, Q)$, a *loss* function on model $\mathcal{P}$, defined to be $\mathcal{B}_d/\mathcal{B}_1$-measurable for every $Q \in \mathcal{P}$. Observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ will be assumed iid (independent and identically distributed). For notational clarity, we shall often write $\lambda_n(Q) := \lambda(\boldsymbol{x}_n, Q)$, not forgetting it is indeed random. Expectations will be taken with respect to the candidate distributions, and our notation is standard; for measurable $g : \mathbb{R}^d \to \mathbb{R}$, we have $\mathbf{E}_Q g := \int g(\boldsymbol{x})\, Q(d\boldsymbol{x})$. In particular we denote the mean and variance of $\boldsymbol{x}$ under Q by $\boldsymbol{\mu}_Q := \int \boldsymbol{x}\, Q(d\boldsymbol{x})$ and $\mathbf{V}_Q := \mathbf{E}_Q(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}_Q \boldsymbol{\mu}_Q^T$ respectively. In working with densities, we assume existence of a dominant measure τ on $\mathcal{B}_d$ where all $Q \ll \tau$ over $\mathcal{P}$, and denote $q := dQ/d\tau$. Useful background references are Halmos [23], Ash and Doleans-Dade [3]. Separate from probability considerations, on an inner product space $\mathcal{X}$ with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \| := |\langle \cdot, \cdot \rangle|^{1/2}$, we use standard notation. The open r-ball at $\boldsymbol{u} \in \mathcal{X}$ is $rB(\boldsymbol{u}) := \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{u}\| < r\}$, for $\mathcal{S} \subset \mathcal{X}$ the closure is $\bar{\mathcal{S}}$, and the interior and boundary are $\text{int}(\mathcal{S}) := \{\boldsymbol{x} : \exists \delta > 0, B_\delta(\boldsymbol{x}) \subset \mathcal{S}\}$ and $\text{bd}(\mathcal{S}) := \bar{\mathcal{S}} \setminus \text{int}(\mathcal{S})$ respectively. With respect to $\mathcal{X}$, $\mathcal{S}$ is open if $\mathcal{S} = \text{int}(\mathcal{S})$ and closed if $\mathcal{S} = \bar{\mathcal{S}}$. Let $\mathbb{S}^d$ denote all symmetric $d \times d$ matrices (here assumed real), with the subset of positive definite matrices denoted $\mathbb{S}_+^d = \{\boldsymbol{A} \in \mathbb{S}^d : \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle > 0, \boldsymbol{x} \neq 0\}$. We may alternatively use $\boldsymbol{A} > 0$ for $\boldsymbol{A} \in \mathbb{S}_+^d$ and $\boldsymbol{A} \geq 0$ for non-negative definite $\boldsymbol{A}$. On $\mathbb{S}^d$ we use the Frobenius norm defined $\|\boldsymbol{A}\| := \sqrt{\text{tr}\, \boldsymbol{A}^T \boldsymbol{A}}$.

Considering arbitrary $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^d$ we see $\|\boldsymbol{A}-\boldsymbol{B}\|^2 = \sum_{i=1}^d (\boldsymbol{a}_i - \boldsymbol{b}_i)^T(\boldsymbol{a}_i - \boldsymbol{b}_i) = \|\boldsymbol{a}-\boldsymbol{b}\|^2$, where $\boldsymbol{a}_j$ denotes the jth column vector of $\boldsymbol{A}$, and $\boldsymbol{a}$ is the $d(d+1)/2$-length vector specifying $\boldsymbol{A}$. That is, the Frobenius norm for $d \times d$ matrices coincides with the Euclidean norm on $\mathbb{R}^{d\times d}$. The ■ and □ symbols respectively denotes the end of examples and proofs, indicating the resumption of the main text.

We assume a fixed model $\mathcal{P}$ on $(\mathbb{R}^d, \mathcal{B}_d)$ for the definitions below, and immediately verify a basic fact.

Definition 1. A loss function λ is said to be *$\mathcal{P}$-proper* if

$$\mathbf{E}_Q \lambda(P) \geq \mathbf{E}_Q \lambda(Q), \quad \forall\, P, Q \in \mathcal{P} \tag{1}$$

and similarly, λ is *strictly $\mathcal{P}$-proper* if (1) holds and

$$\mathbf{E}_Q \lambda(P) = \mathbf{E}_Q \lambda(Q) \iff P = Q. \tag{2}$$

Letting $\Theta \subset \mathbb{R}^m$ for some $m > 0$, for our purposes, we call $(\mathcal{P}, v, \Theta)$ a *parametric model* when $v : \mathcal{P} \to \Theta$, called a *parametrization* of $\mathcal{P}$, is surjective. Denoting $v^{-1}(\boldsymbol{\theta}) \coloneqq \{Q \in \mathcal{P} : v(Q) = \boldsymbol{\theta}\}$, if we consider analogous loss $L(\boldsymbol{\theta}) \coloneqq L(\boldsymbol{x}, \boldsymbol{\theta})$ defined for $\boldsymbol{\theta} \in \Theta$, we say that L is *Θ-proper* if

$$\mathbf{E}_Q L(\boldsymbol{\theta}') \geq \mathbf{E}_Q L(\boldsymbol{\theta}), \quad \forall\, Q \in v^{-1}(\boldsymbol{\theta}) \tag{3}$$

holds for every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$. We say L is *strictly Θ-proper* if

$$\text{(3) holds with equality} \iff \boldsymbol{\theta} = \boldsymbol{\theta}'. \tag{4}$$

Proposition 2. *Given a parametric model $(\mathcal{P}, v, \Theta)$, loss $L(\boldsymbol{\theta})$, and defining $\lambda_v(Q) \coloneqq L(v(Q))$ on $Q \in \mathcal{P}$, we have*

$$L \text{ is } \Theta\text{-proper} \iff \lambda_v \text{ is } \mathcal{P}\text{-proper}.$$

If v is bijective, the equivalence holds analogously for strict $\Theta/\mathcal{P}$ propriety.

Proof. For notational clarity, we denote $\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta}) \coloneqq L(\cdot, \boldsymbol{\theta}') - L(\cdot, \boldsymbol{\theta})$ and similarly $\Delta_{\lambda_v}(P, Q) \coloneqq \lambda_v(\cdot, P) - \lambda_v(\cdot, Q)$. By the surjectivity of v, $v^{-1}(\boldsymbol{\theta}) \neq \emptyset$ on Θ. Let L be Θ-proper. Then

$$\begin{aligned} 0 &\leq \boldsymbol{E}_Q[L(\cdot, v(P)) - L(\cdot, v(Q))], \quad \forall\, P, Q \in \mathcal{P} \\ &= \boldsymbol{E}_Q[\lambda_v(\cdot, P) - \lambda_v(\cdot, Q)], \end{aligned}$$

where the latter equality follows by definition of λ_v.

For the other direction, first note that

$$\boldsymbol{E}_R[L(\cdot, v(R))] = \boldsymbol{E}_R[L(\cdot, v(Q))] = \boldsymbol{E}_R[L(\cdot, \boldsymbol{\theta})],$$

for all $R \in v^{-1}(\boldsymbol{\theta})$. Take arbitrary $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$. For some $P \in v^{-1}(\boldsymbol{\theta}')$ and $Q, R \in v^{-1}(\boldsymbol{\theta})$, we have

$$\begin{aligned} \boldsymbol{E}_R[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] &= \boldsymbol{E}_R[L(\cdot, v(P))] - \boldsymbol{E}_R[L(\cdot, v(Q))] \\ &= \boldsymbol{E}_R[\lambda_v(\cdot, P)] - \boldsymbol{E}_R[\lambda_v(\cdot, R)] \\ &\geq 0, \end{aligned}$$

where the final inequality follows by $\mathcal{P}$-propriety.

Now for the strict case. Let v be bijective, and assume strict $\mathcal{P}$-propriety of $\lambda(v)$. Note that $\boldsymbol{\theta} = \boldsymbol{\theta}'$ trivially implies $\boldsymbol{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] = 0$. Conversely, assume $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$. As $v^{-1}(\boldsymbol{\theta}) \cap v^{-1}(\boldsymbol{\theta}') = \emptyset$, we have for $P \in v^{-1}(\boldsymbol{\theta}')$ and $Q \in v^{-1}(\boldsymbol{\theta})$ that $P \neq Q$. Then by strict $\mathcal{P}$-propriety $0 \neq \boldsymbol{E}_Q[\Delta_{\lambda_v}(P, Q)] = \boldsymbol{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})]$, which is to say there exists $Q \in v^{-1}(\boldsymbol{\theta})$ such that the expected difference does not vanish. The sufficiency result follows by contrapositive.

Next let strict Θ-propriety of L hold. That $P = Q$ implies the expectation of the difference vanishes is immediate. Conversely, let $P \neq Q$. Then injectivity of v implies $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, and as such $\boldsymbol{E}_Q[\Delta_{\lambda_v}(P, Q)] = \boldsymbol{E}_Q[\Delta_L(\boldsymbol{\theta}', \boldsymbol{\theta})] \neq 0$ using strict Θ-propriety, which gives us the result. □

Some remarks are in order before we proceed. Given the parametric models considered here, for a given model and loss $L(\boldsymbol{\theta})$ indexed by a subset of Euclidean space, L generates a loss λ_v on $\mathcal{P}$, whose propriety is equivalent to that of Θ propriety of L. Recall that we seek a systematic approach to derive $\mathcal{P}$-proper losses. Working with functions on Euclidean space is intuitive and analytically much simpler, and thus instead of working directly on $\mathcal{P}$, considering the statements in Prop. 2, it will be natural to derive a (strictly) proper λ_v by first constructing a (strictly) proper L. The parametrization defined here is sufficient for our arguments, though a more general treatment would require a different sort of parametrization map, from Θ to L_2, with regularity conditions including Fréchet differentiability. A comprehensive advanced reference for interested readers is Bickel et al. [7].

As introduced in Section 1, we are assuming an estimation task using the λ-optimizing estimator $\widehat{P}_N := \arg\min_{Q \in \mathcal{P}} \sum_{n=1}^{N} \lambda_n(Q)$ given sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. We note that $\widehat{P}_N$ has a natural interpretation as a minimum divergence estimator which minimizes the distance from the unknown underlying distribution. We call $d : \mathcal{P}^2 \to \mathbb{R}$ a *divergence* or *contrast* on $\mathcal{P}$ when $d \geq 0$ and $d(P, Q) = 0 \iff P = Q$. Define $d_\lambda(P, Q) := \mathbf{E}_Q(\lambda(P) - \lambda(Q))$, and assume $\mathbf{E}_Q|\lambda(P)| < \infty$ for all $P, Q \in \mathcal{P}$. Then the strong law of large numbers gives us that

$$N^{-1} \sum_{n=1}^{N} (\lambda_n(P) - \lambda_n(P_X)) \to d_\lambda(P, P_X)$$

almost surely, and trivially, any estimate Q such that $\mathbf{E}_{P_X} \lambda(Q)$ is minimal equivalently ensures $d_\lambda(Q, P_X)$ is minimal. More importantly, if we can show the $\mathcal{P}$-propriety of λ, the non-negativity of d_λ follows, and similarly strict $\mathcal{P}$ propriety of λ implies that d_λ is indeed a valid divergence. When $\mathcal{P}$ and Θ are isomorphic, strict propriety of losses is equivalent across the two domains. In such a case, if we have an approach to derive proper losses for a given class of models, then we equivalently have an approach to derive valid divergences on the same models. Such an approach will be introduced in Section 2.2.

2.2 Deriving estimators using proper loss functions

The basic idea is to construct L on Euclidean space, show (strict) Θ-propriety, and thus given an appropriate v then show (strict) $\mathcal{P}$-propriety of the loss λ_v generated by L. What makes this approach particularly fruitful is the fact that we can make liberal use of fundamental properties of convex functions, properties which as we will see are suggestive of an intuitive functional form for $L(\boldsymbol{\theta})$ to take, and which conveniently admit simple proofs. An authoritative classic reference is Rockafellar [41].

Let f be a real concave function on $\mathcal{S} \subset \mathbb{R}^m$. Let us call $\boldsymbol{x}^* \in \mathcal{S}$ a *supergradient* of f at $\boldsymbol{x}$ if $-\boldsymbol{x}^*$ is a subgradient of convex $-f$ at $\boldsymbol{x}$. It then follows that

$$f(\boldsymbol{z}) \leq f(\boldsymbol{x}) + \langle \boldsymbol{x}^*, \boldsymbol{z} - \boldsymbol{x} \rangle, \quad \forall\, \boldsymbol{z} \in \mathcal{S}. \tag{5}$$

In general, a supergradient of f at $\boldsymbol{x}$ need not exist, and should it exist it need not be unique. Denoting the set of all supergradients of f at $\boldsymbol{x}$ by $\partial f(\boldsymbol{x})$, if $\partial f(\boldsymbol{x}) \neq \emptyset$, then the existence of a supergradient $\boldsymbol{x}^* \in \partial f(\boldsymbol{x})$ implies a hyperplane H in $\mathbb{R}^{m+1}$ where

$$H(\boldsymbol{x}) = \{(\boldsymbol{z}, z_0) : (\boldsymbol{z}, z_0)^T(\boldsymbol{x}^*, -1) = \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle - f(\boldsymbol{x})\},$$

recalling f is real-valued on $\mathcal{S}$. The epigraph of concave f is $\text{epi}(f) = \{(\boldsymbol{x}, \alpha) \in S \times \mathbb{R} : \alpha \leq f(\boldsymbol{x})\}$, a convex set. Note for arbitrary $(\boldsymbol{z}, \alpha) \in \text{epi}(f)$, we have that

$$(\boldsymbol{z}, \alpha)^T(\boldsymbol{x}^*, -1) \geq \boldsymbol{z}^T\boldsymbol{x}^* + f(\boldsymbol{z}) \geq \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle - f(\boldsymbol{x})$$

giving us that $\text{epi}(f)$ is a subset of the upper half-space corresponding to $H(\boldsymbol{x})$ whenever $\partial f(\boldsymbol{x})$ is non-empty. Clearly $(\boldsymbol{x}, f(\boldsymbol{x})) \in \text{epi}(f) \cap H(\boldsymbol{x})$, thus the intersection of the epigraph of f and the boundary of the half-space containing it is non-empty, and H is a supporting hyperplane to $\text{epi}(f)$. By Rockafellar [41], we have that if we restrict ourselves to $\boldsymbol{x} \in \text{int}(\text{dom}(f))$ only, then $\partial f(\boldsymbol{x}) \neq \emptyset$ always holds (Thm. 23.4). Furthermore, uniqueness of the supergradient, namely $\partial f(\boldsymbol{x}) = \{\boldsymbol{x}^*\}$ holds if and only if $\boldsymbol{x}^* = \nabla f(\boldsymbol{x})$ (Thm. 25.1), the gradient of f at $\boldsymbol{x}$. Thus, if we consider concave f on an open domain, the supergradient exists everywhere, and it is unique (and coincides with the gradient) if and only if f is differentiable. Designing losses as functions of concave functions and their supergradients allows us to use the inequality (5) for expediting proofs of Θ-propriety. Before giving the general form, we consider a motivating example from Grünwald and Dawid [22] for the discrete case.

Example 3 (Proper losses and Bregman divergences in the discrete case). Let us consider a sample space of $\{1, 2, \ldots, M\}$ and thus specify probability mass functions by $\boldsymbol{p} = (p_1, \ldots, p_M) \in \mathbb{P}_M$, the probability simplex on $\mathbb{R}^M$. Let $\boldsymbol{e}(m) = (I_1(m), \ldots, I_M(m))$, a vector of indicator functions defined $I_k(m) = 1$ if $m = k$, else 0. Then, define $L(m, \boldsymbol{p}) \coloneqq f(\boldsymbol{p}) + \langle \boldsymbol{p}^*, \boldsymbol{e}(m) - \boldsymbol{p} \rangle$, where f is any function concave on $\text{int}(\mathbb{P}_M)$, and $\boldsymbol{p}^* \in \partial f(\boldsymbol{p})$. Since $\mathbf{E}_q \boldsymbol{e} = \boldsymbol{q}$, we may note that

$$\mathbf{E}_q(L(m, \boldsymbol{p}) - L(m, \boldsymbol{q})) = f(\boldsymbol{p}) - f(\boldsymbol{q}) + \langle \boldsymbol{p}^*, \boldsymbol{q} - \boldsymbol{p} \rangle \geq 0$$

for all $\boldsymbol{p}, \boldsymbol{q} \in \text{int}(\mathbb{P}_M)$, which follows from (5). That is, L is $\text{int}(\mathbb{P}_M)$-proper. Without loss of generality we may consider $\mathbb{P}_M$ to be a subset of $\mathbb{R}^{M-1}$ equal to $\{\boldsymbol{p} \in \mathbb{R}^{M-1} : \sum_{i=1}^{M-1} p_i \leq 1, p_i \geq 0\}$, and thus $\text{int}(\mathbb{P}_M)$ is all $\boldsymbol{p} \in \mathbb{P}_M$ such that $\sum_{i=1}^{M-1} p_i < 1$ and $p_i > 0, i = 1, \ldots, M$. This implies that the Mth coordinate of any $\boldsymbol{p} \in \text{int}(\mathbb{P}_M)$ is in $(0, 1)$. The set of all discrete probability distributions on sample space $\{1, 2, \ldots, M\}$ assigning non-zero probability to every event is clearly isomorphic to $\text{int}(\mathbb{P}_M)$ and thus L generates a proper loss (strictly proper if f is strictly concave) on that space by Prop. 2. Recalling the close relation between proper scoring rules and distance-like metrics (Section 2.1), one may also note that with some additional regularity conditions on f, including its differentiability on $\text{int}(\mathbb{P}_M)$, then we would have

$$\begin{aligned}
\mathbf{E}_q(L(m, \boldsymbol{p}) - L(\cdot, \boldsymbol{q})) &= f(\boldsymbol{p}) - f(\boldsymbol{q}) + \langle \nabla f(\boldsymbol{p}), \boldsymbol{q} - \boldsymbol{p} \rangle \\
&= (-f)(\boldsymbol{q}) - (-f)(\boldsymbol{p}) - \langle \nabla(-f)(\boldsymbol{p}), \boldsymbol{q} - \boldsymbol{p} \rangle \\
&= d_B(\boldsymbol{q}, \boldsymbol{p}; -f)
\end{aligned}$$

where $d_B(\cdot, \cdot; -f)$ is none other than the Bregman divergence with respect to convex function $-f$. The literature on Bregman divergences is rich, with particularly interesting research from the statistics and machine learning communities [4]. Finally, note similar constructions yielding Bregman divergences in the continuous case also exist [22]. ■

The above example suggests the utility of working with loss functions composed of (strictly) concave functions and their supergradients, namely in terms of verifying (strict) propriety. We now return to our original setting and give sufficient conditions for propriety of losses taking a particular functional form. It is readily observed that this form is a straightforward generalization of the example above.

Proposition 4. *Let $(\mathcal{P}, \Theta, v)$ be a parametric model, where Θ is an open subset of $\mathbb{R}^k$. Consider the family of loss functions of the form*

$$L(\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}^*, h(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{\theta} \rangle, \quad \boldsymbol{\theta} \in \Theta \tag{6}$$

where f on Θ is concave and $h(\cdot, \boldsymbol{\theta})$ is $\mathcal{B}_d/\mathcal{B}_k$-measurable. Then, if there exists an operator $\boldsymbol{\theta} \mapsto r(\boldsymbol{\theta})$ returning values in Θ such that for all $Q \in \mathcal{P}$,

(i) $\mathbf{E}_Q h(\boldsymbol{\theta}) = v(Q) + r_Q(\boldsymbol{\theta})$

(ii) $\langle \boldsymbol{\theta}^*, r(\boldsymbol{\theta}) \rangle \geq 0, \forall\, \boldsymbol{\theta} \in \Theta$, *and* $r(v(Q)) = 0$

noting r can depend on Q, then we have that λ_v, the loss generated by L given $(\mathcal{P}, \Theta, v)$, is a $\mathcal{P}$-proper loss.

Proof. As discussed at the start of Section 2.2, the existence of supergradient $\boldsymbol{\theta}^*$ at $\boldsymbol{\theta}$ follows as $\operatorname{int}(\operatorname{dom}(f)) = \operatorname{dom}(f)$ by openness of Θ. Fix $\boldsymbol{\theta}$ and $Q \in v^{-1}(\boldsymbol{\theta})$. Noting we have $\mathbf{E}_Q L(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$, and so for arbitrary $\boldsymbol{\theta}' \in \Theta$, we have that

$$\mathbf{E}_Q(L(\boldsymbol{\theta}') - L(\boldsymbol{\theta})) = f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}'^*, \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle + \langle \boldsymbol{\theta}'^*, r(\boldsymbol{\theta}') \rangle.$$

By (ii), the last term on the right-hand side is non-negative. The sum of the first three terms is non-negative by assumptions on f and inequality (5). □

Corollary 5. *Given the assumptions of Prop. 4, if f is strictly convex, then the λ_v generated by L in (6) is strictly $\mathcal{P}$-proper.*

Proof. Fixing $\boldsymbol{\theta}$ and $Q \in v^{-1}(\boldsymbol{\theta})$, since $r(\boldsymbol{\theta}')$ vanishes when $\boldsymbol{\theta}' = \boldsymbol{\theta}$, it remains only to show that definiteness of the first three terms. This follows readily from strict concavity (see Appendix). □

In an inference problem, either due to *a priori* knowledge, or for reasons of ease of interpretation, simplicity, and computational and analytical tractability, assumptions of parametric statistical models are often made. That is to say, we begin with some $(\mathcal{P}, v, \Theta)$ as defined above. The easiest situation is when we already have L known to be of the form given in (6). Then, using the conditions given in Prop. 4, in many cases a direct examination of $\mathbf{E}h(\boldsymbol{\theta})$ will be all that is required. Things become slightly less immediate when the task is to construct a new loss, though in that case one naturally would begin with a concave f, and select an appropriate h given the model and knowledge of the supergradient, from which the desired propriety would follow.

The remaining results essentially deal with characterizing combinations of important model and loss classes for which the verification of propriety is, given the results, all but immediate. Before stating these results however, we give a series of examples illustrating the selection of a rudimentary h for propriety proofs in a number of basic cases.

Example 6 (Proper losses: identification by mean). For many models of interest, it will be that for some fixed open subset $\Theta_1 \subset \mathbb{R}^d$, every P is such that $\boldsymbol{\mu}_P \coloneqq \mathbf{E}_P \boldsymbol{x} \in \Theta_1$. Making this more concrete, if we define $\mathcal{P}_1 \coloneqq \{P : \boldsymbol{\mu}_P \in \Theta_1\}$, then letting $v_1(P) \coloneqq \boldsymbol{\mu}_P$, obviously $(\mathcal{P}_1, v_1, \Theta_1)$ is a parametric model. Let f_1 be a strictly concave function on Θ_1. Then, defining $L_1(\boldsymbol{x}, \boldsymbol{\theta}) \coloneqq f_1(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}^*, \boldsymbol{x} - \boldsymbol{\theta} \rangle$, we note that this takes the form given in (6), where $h(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}$ for all $\boldsymbol{\theta} \in \Theta$. Clearly, as $\mathbf{E}_P h(\boldsymbol{x}, \boldsymbol{\theta}) = v_1(P)$ (and thus $r_P = 0$), by Prop. 4, we have that λ_{v_1} generated by L_1 is a $\mathcal{P}_1$-proper loss, and a strictly proper loss on the set of equivalence classes on $\mathcal{P}_1$ identified by their mean. For example, say $\Theta_1 = (0, \infty)$ and $f_1(\cdot) = \log(\cdot)$. Then $L_1(x, \theta) = \log \theta + (x - \theta)/\theta$ has the noted properties. ■

Example 7 (Proper losses: identification by second moment). One may easily complete an analogous exercise for the second moments of $\boldsymbol{x}$. First note that $\mathbb{S}^d$ is a subset of $\mathbb{R}^{d \times d}$ and is clearly isomorphic to $\mathbb{R}^{d(d+1)/2}$, since given any fixed rule for building matrices from long vectors, every vector $\boldsymbol{a} \in \mathbb{R}^{d(d+1)/2}$ specifies one and only one $\boldsymbol{A} \in \mathbb{S}^d$. Fixing some open $\Theta_2 \subset \mathbb{S}^d$, and $\mathcal{P}_2 \coloneqq \{P : \mathbf{E}_P \boldsymbol{x}\boldsymbol{x}^T \in \Theta_2\}$, for strictly concave f_2, and $v_2(P) \coloneqq \mathbf{E}_P \boldsymbol{x}\boldsymbol{x}^T$, we have by an identical argument that $L_2(\boldsymbol{x}, \boldsymbol{W}) \coloneqq f_2(\boldsymbol{W}) + \langle \boldsymbol{W}^*, \boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{W} \rangle$ generates a $\mathcal{P}_2$-proper loss λ_{v_2}, which is strictly proper on the analogous set of equivalence classes. We may readily note here that $h(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{x}\boldsymbol{x}^T$, so $\mathbf{E}_P h(\boldsymbol{x}, \boldsymbol{W}) = v_2(P)$ for all $\boldsymbol{W} \in \Theta_2$, and again $r_P = 0$. ■

Members of the family of d-dimensional Gaussian distributions are identified by mean and variance. The following example includes this situation as an important special case.

Example 8 (Proper losses: identification by variance). Let us begin by assuming the mean $\bar{\boldsymbol{x}}$ is known, and define a general model $\mathcal{P}_3 \coloneqq \{P : \boldsymbol{\mu}_P = \bar{\boldsymbol{x}}, \mathbf{V}_P \in \Theta_3\}$, where Θ_3 is an open subset of $\mathbb{S}^d$. Naturally set the parametrization to $v_3(P) \coloneqq (\mathbf{V}_P)$. Since $\bar{\boldsymbol{x}}$ is fixed and known, we may use it in defining a given h, and intuitively set $h(\boldsymbol{x}, \boldsymbol{W}) \coloneqq (\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^T$. One may readily observe

$$\mathbf{E}_P h(\boldsymbol{x}, \boldsymbol{W}) = \mathbf{V}_P + (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_P)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_P)^T = v_3(P)$$

since the second term, namely $r(\boldsymbol{W})$, is zero for all P by definition of $\mathcal{P}_3$. For strictly concave f_3 on Θ_3 then, using Prop. 4 we immediately have that $L_3(\boldsymbol{x}, \boldsymbol{W}) \coloneqq f_3(\boldsymbol{W}) + \langle \boldsymbol{W}^*, (\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^T - \boldsymbol{W} \rangle$ generates a $\mathcal{P}_3$-proper loss function λ_{v_3}, which is strictly proper on the set of corresponding equivalence classes. ■

The loss L_3 considered in the example above takes on a form which is intuitively appealing, but the propriety proof above assumed a fixed, known mean. What if the mean and variance are both unknown, and we seek to estimate them both? Do losses of this form retain propriety? Under slightly stronger assumptions, we indeed may retain propriety using losses of the exact same form, without requiring the mean to be known. We prove this directly below, without appealing to Prop. 4. Note that we call a function g on $\mathbb{S}^d$ non-decreasing if $\boldsymbol{A} \geq \boldsymbol{B} \implies g(\boldsymbol{A}) \geq g(\boldsymbol{B})$, where we say $\boldsymbol{A} \geq \boldsymbol{B}$ when for $\boldsymbol{A}, \boldsymbol{B} \geq 0$ we have $\boldsymbol{A} - \boldsymbol{B} \geq 0$. Increasing functions are defined analogously.

Proposition 9. *Again consider random vector $\boldsymbol{x} \in \mathbb{R}^d$ and parameter space $\Theta \coloneqq \Theta_m \times \Theta_v \subset \mathbb{R}^d \times \mathbb{S}_+^d$ where Θ_v is assumed open. Denote each $\boldsymbol{\theta} = (\boldsymbol{u}, \boldsymbol{W})$, let f be concave and non-decreasing on Θ_v, define parametrization $v(P) \coloneqq (\boldsymbol{\mu}_P, \mathbf{V}_P)$, and let $\mathcal{P}$ be such that $(\mathcal{P}, v, \Theta)$ is a parametric model. Then, the loss function*

$$L(\boldsymbol{x}, \boldsymbol{\theta}) \coloneqq f(\boldsymbol{W}) + \langle \boldsymbol{W}^*, (\boldsymbol{x} - \boldsymbol{u})(\boldsymbol{x} - \boldsymbol{u})^T - \boldsymbol{W} \rangle \tag{7}$$

generates a $\mathcal{P}$-proper loss λ_v. If f is strictly concave, increasing, and v is bijective, then propriety holds strictly.

Proof. First note $\langle \boldsymbol{W}, \boldsymbol{x}\boldsymbol{x}^T \rangle = \operatorname{tr} \boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^T = \boldsymbol{x}^T\boldsymbol{W}\boldsymbol{x}$ for $\boldsymbol{W} \in \mathbb{S}^d$, and so $\mathbf{E}_Q\langle \boldsymbol{W}^*, (\boldsymbol{x}-\boldsymbol{u})(\boldsymbol{x}-\boldsymbol{u})^T \rangle = \langle \boldsymbol{W}^*, \mathbf{E}_Q(\boldsymbol{x}-\boldsymbol{u})(\boldsymbol{x}-\boldsymbol{u})^T \rangle$. Fixing $\boldsymbol{\theta} = (\boldsymbol{u}, \boldsymbol{W}) \in \Theta$ and $Q \in v^{-1}(\boldsymbol{\theta})$, some algebra readily yields

$$\mathbf{E}_Q L(\boldsymbol{x}, (\boldsymbol{u}', \boldsymbol{W}')) = f(\boldsymbol{W}) - \langle \boldsymbol{W}'^*, \boldsymbol{W}' \rangle + \langle \boldsymbol{W}'^*, \mathbf{V}_Q + (\boldsymbol{u}' - \boldsymbol{\mu}_Q)(\boldsymbol{u}' - \boldsymbol{\mu}_Q)^T \rangle$$

for arbitrary $\boldsymbol{\theta}' = (\boldsymbol{u}', \boldsymbol{W}')$. Clearly, $\mathbf{E}_Q L(\boldsymbol{x}, (\boldsymbol{u}, \boldsymbol{W})) = f(\boldsymbol{W})$, and thus the expected difference is

$$\begin{aligned}\mathbf{E}_Q(L(\boldsymbol{\theta}') - L(\boldsymbol{\theta})) = f(\boldsymbol{W}') - f(\boldsymbol{W}) \\ + \langle \boldsymbol{W}'^*, \mathbf{V}_Q - \boldsymbol{W}' \rangle + \langle \boldsymbol{W}'^*, (\boldsymbol{u}' - \boldsymbol{\mu}_Q)(\boldsymbol{u}' - \boldsymbol{\mu}_Q)^T \rangle.\end{aligned}$$

The sum of the first three terms is non-negative by inequality (5), and definiteness follows if f is strictly concave. In this case, the first three terms vanish if and only if $\boldsymbol{W}' = \boldsymbol{W}$. The last term is equal to $(\boldsymbol{u}' - \boldsymbol{\mu}_Q)^T \boldsymbol{W}'^* (\boldsymbol{u}' - \boldsymbol{\mu}_Q)$. As f is non-decreasing, $\boldsymbol{W}'^* \geq 0$, thus making the term non-negative. If f is increasing, we have $\boldsymbol{W}'^* > 0$ and thus vanishes if and only if $\boldsymbol{u}' = \boldsymbol{u}$. Non-negativity of the expected difference thus holds under weak assumptions for all $\boldsymbol{\theta}'$, and under the stronger assumptions the difference is zero if and only if $\boldsymbol{\theta}' = \boldsymbol{\theta}$. The result then follows from Prop. 2. □

The above result implies a natural class of loss functions, as is summarized in the following Corollary.

Corollary 10. *Let g be a non-decreasing, concave real function on $(0, \infty)$. Then, given any parametric model $(\mathcal{P}, v, \Theta)$ satisfying the assumptions of Prop. 9, we have that any*

$$L(\boldsymbol{x}, \boldsymbol{\theta}) \coloneqq g((\det \boldsymbol{W})^{1/d}) + \langle \boldsymbol{W}^*, (\boldsymbol{x}-\boldsymbol{u})(\boldsymbol{x}-\boldsymbol{u})^T - \boldsymbol{W} \rangle \tag{8}$$

satisfies the same propriety properties as any loss of the form (7).

Proof. Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$ for integer $m > 1$, $\alpha \in [0, 1]$, and let g be non-decreasing, concave real function on $(0, \infty)$, while f is assumed real and concave on $\mathbb{R}^m$. Then by our assumptions,

$$\begin{aligned}g(f(\alpha\boldsymbol{a} + (1-\alpha)\boldsymbol{b})) &\geq g(\alpha f(\boldsymbol{a}) + (1-\alpha) f(\boldsymbol{b})) \\ &\geq \alpha g(f(\boldsymbol{a})) + (1-\alpha) g(f(\boldsymbol{b}))\end{aligned}$$

that is $g \circ f$ is concave. If f and g are strictly concave, or f is strictly concave and g is increasing, then the composition is strictly concave.

Next, note that for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^d_+$, we have

$$\begin{aligned}(\det(\alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{B}))^{1/d} &\geq (\det \alpha\boldsymbol{A})^{1/d} + (\det(1-\alpha)\boldsymbol{B})^{1/d} \\ &= \alpha(\det \boldsymbol{A})^{1/d} + (1-\alpha)(\det \boldsymbol{B})^{1/d}\end{aligned}$$

where the first inequality is Minkowski's determinant theorem. This gives us concavity of $(\det \boldsymbol{W})^{1/d}$, and the composition results above alongside Prop. 9 yield the desired result. □

A few additional remarks should be made here. The above Corollary has us using the proof of propriety of a rudimentary loss function to trivially obtain propriety results for a much larger class of losses. This is possible in general when we only use Prop. 2 in proving the rudimentary case (i.e., Prop. 9), but need not hold when propriety is verified using the conditions given by Prop. 4. The reason for this is that even with a fixed parametric model, the $h(\cdot, \boldsymbol{\theta})$ of course may depend on f in such a way that when we replace f with some concave composition $g \circ f$, the same h need not satisfy condition (ii) any longer. In the following example, we give concrete examples of losses taking the form (8).

Example 11 (Variance-normalized proper loss functions). Given the same assumptions as Prop. 9, we consider two obvious choices for g in (8), namely $\log(\cdot)$ and $(\cdot)^{1/\alpha}$ for real $\alpha \geq 1$. Both are clearly concave and non-decreasing on $(0, \infty)$. Differentiability will also be of use since the resulting supergradients will coincide with the gradients. Noting that on $\mathbb{S}_+^d$, have $\partial_{i,j} \det \boldsymbol{W} := \partial \det \boldsymbol{W} / \partial w_{i,j} = (\det \boldsymbol{W}) \operatorname{tr} \boldsymbol{W}^{-1} \partial \boldsymbol{W} / \partial w_{i,j}$, and the matrix which has the partial derivative $\partial_{i,j} \det \boldsymbol{W}$ as its (i,j)th element is readily confirmed to be $\partial \det \boldsymbol{W} / \partial \boldsymbol{W} := [\partial_{i,j} \det \boldsymbol{W}] = \det(\boldsymbol{W}) \boldsymbol{W}^{-1}$. Let g in (8) be replaced by $(\cdot)^{1/\alpha}$ (for $\alpha \geq d$). With $\boldsymbol{\theta} = (\boldsymbol{u}, \boldsymbol{W})$, the resulting loss is $L_\alpha(\boldsymbol{x}, \boldsymbol{\theta}) = (\det \boldsymbol{W})^{1/\alpha}(1 + (\boldsymbol{x} - \boldsymbol{u})^T \boldsymbol{W}^{-1}(\boldsymbol{x} - \boldsymbol{u})/\alpha)$ and it generates loss

$$\lambda_v^\alpha(\boldsymbol{x}, P) = (\det \mathbf{V}_P)^{1/\alpha} \left(1 + \frac{(\boldsymbol{x} - \boldsymbol{\mu}_P)^T \mathbf{V}_P^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_P)}{\alpha}\right). \tag{9}$$

Similarly, letting g be $\log(\cdot)$, the analogous L_D generates loss

$$\lambda_v^D(\boldsymbol{x}, P) = \log \det \mathbf{V}_P + (\boldsymbol{x} - \boldsymbol{\mu}_P)^T \mathbf{V}_P^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_P). \tag{10}$$

As we used the form in Corollary 10 as-is, it immediately follows that on *any* parametric model $(\mathcal{P}, v, \Theta)$, $\Theta \subset \mathbb{R}^d \times \mathbb{S}_+^d$, λ_v^α and λ_v^D are $\mathcal{P}$-proper. λ_v^D is strictly $\mathcal{P}$-proper whenever v is bijective (see Appendix). ■

The results and examples above have all been fairly general, but propriety for many well-known model/loss combinations are trivial consequences of the facts verified thus far. In Section 2.3, we consider some specific cases.

2.3 Application to common parametric models

Fix a random vector $\boldsymbol{x}$ as in Section 2.2, and for concreteness consider λ_v^D given in (10). Clearly, λ_v^D is proper on the set of all Q such that variance $\mathbf{V}_Q$ of $\boldsymbol{x}$ is finite. Thus, we have a proper loss function for simultaneous estimation of all the parameters of models including the d-dimensional Normal, truncated Normal, Gamma, Weibull, generalized extreme value, and log-Normal, among many others (identical statements hold for L_α). A particularly useful fact is that propriety of λ_v^D holds strictly for the d-dimensional Normal case. If we treat particular parameters as given, then λ_v^D is strictly proper for many more distribution families. For example, fixing the shape parameter of a Gamma or Weibull, we have strict propriety of λ_v^D on the set of all Gamma/Weibull distributions sharing that shape. Analogous results hold for different parameters. All of these facts follow immediately from Corollary 10 and the discussion in the example ending Section 2.2.

Evaluating λ_v^D and λ_v^α is easy for the distributions mentioned. As an illustrative example, we look at λ_v^α for the Weibull distribution, defined by distribution function $W(x; \xi, \kappa) := 1 - \exp(-(x/\xi)^\kappa)$ with density

$$w(x; \xi, \kappa) = \frac{\kappa}{\xi} \left(\frac{x}{\xi}\right)^{\kappa - 1} \exp\left(-\left(\frac{x}{\xi}\right)^\kappa\right).$$

The two-parameter Weibull is a family of continuous probability distributions which arises naturally in the theory of extreme values. It has seen extensive use by practitioners for modelling systems as diverse as survival rates, resource allocation, and wind velocity [40]. Given scalar observations $x_1, \ldots, x_N$ we naturally approximate the λ_v^α-optimal estimator by $(\widehat{\xi}_N, \widehat{\kappa}_N)_\alpha := \arg\min_{\xi,\kappa} \sum_{n=1}^N \lambda_v^\alpha(x_n, (\xi, \kappa))$, where from (9) the exact form of the loss is

$$\lambda_v^\alpha(x_n, (\xi, \kappa)) = \left(\xi^2(\Gamma_2 - \Gamma_1^2)\right)^{1/\alpha} \left(1 + \frac{(x_n - \xi\Gamma_1)^2}{\alpha \xi^2 (\Gamma_2 - \Gamma_1^2)}\right)$$

where $\Gamma_m \coloneqq \Gamma(1 + m/\kappa)$ and $\Gamma(\cdot)$ denotes the usual Gamma function [2]. For reference, we include the explicit form of the first-order derivatives. Denoting $\lambda_v^\alpha \coloneqq \lambda_v^\alpha(x, (\xi, \kappa))$, we have that

$$\frac{\partial \lambda_v^\alpha}{\partial \xi} = \frac{2(\xi^{2-\alpha} G)^{1/\alpha}}{\alpha} \left(1 - \frac{1}{\alpha} - \frac{1}{\xi G}\left((x - \xi\Gamma_1)\Gamma_1 + \left(1 - \frac{1}{\alpha}\right)\frac{(x - \xi\Gamma_1)^2}{\xi}\right)\right)$$

$$\frac{\partial \lambda_v^\alpha}{\partial \kappa} = \frac{2}{\alpha\kappa^2}\left(\frac{\xi^2}{G^{\alpha-1}}\right)^{1/\alpha}\left(\left(1 - \frac{1}{\alpha}\right)(\Gamma_2' - \Gamma_1\Gamma_1')\left(\frac{(x - \xi\Gamma_1)^2}{\xi^2 G} - 1\right) + \frac{(x - \xi\Gamma_1)\Gamma_1'}{\xi}\right)$$

where we denote $G \coloneqq (\Gamma_2 - \Gamma_1^2)$ and $\Gamma_m' \coloneqq \Gamma'(1 + m/\kappa)$. Of course similar expressions may be found just as easily for $\lambda_v^D(x, (\xi, \kappa))$.

3 Experimental results

Here we present the results of empirical tests conducted to evaluate the utility of representative proper losses as derived in Section 2, in the context of parameter estimation. Tests include controlled simulations as well as estimation tasks using real-world data. Direct comparisons are made between the estimation accuracy of minimum proper loss estimators given above, and several standard minimum loss/divergence estimators from the literature. Basic results for preliminary work were reported in Holland and Ikeda [25].

3.1 Controlled estimation task

Before we deal with real-world data where the underlying distribution is inherently unknown, we begin with some controlled tests. The task is scale parameter estimation given iid data, respectively for Gaussian, Gamma, and Weibull distributions, assuming shape/shift parameters are known. Sample size per trial for each distribution is 500, and 10 independent trials are carried out, with mean absolute estimate error computed over these trials. Gamma and Weibull shape is fixed at 2, and the true scale parameter for all three cases is 1.772454. Results are shown in Fig. 1.

The estimation methods used are methods of moments (MOM), minimum negative log likelihood (NLL), minimum λ_v^α estimation (DET, with $\alpha = 1.01$), minimum λ_v^D estimation (LOGDET), minimum Hellinger distance (HELLI), and minimum total variation distance (TVD). MOM for Gamma and Gaussian cases follows the classical approach. For Weibull MOM, we use the method of Newby [33]. All NLL implementations use the `optimize` routine in the R language and environment [39], as do DET/LOGDET optimizations in this simple univariate case. HELLI and TVD use the R library `distrEx` and `optimize` to minimize the dissimilarity between the candidate and empirical densities, with the latter symmetrized. The average TVD error in the Gamma case was much larger than the others (> 5.00) and thus to better illustrate relative performance we have removed it from the bar plot. In this simple case, we see that the minimum proper loss estimators put forward in Section 2.2 are competitive with all alternative methods considered here.

3.2 Predictive parameter estimation task

As an interesting and challenging application of the methods discussed in Section 2, we take an in-depth look at a probabilistic forecasting problem using real-world data. Here we give an overview of the data and the basic problem formulation, introduce several standard reference

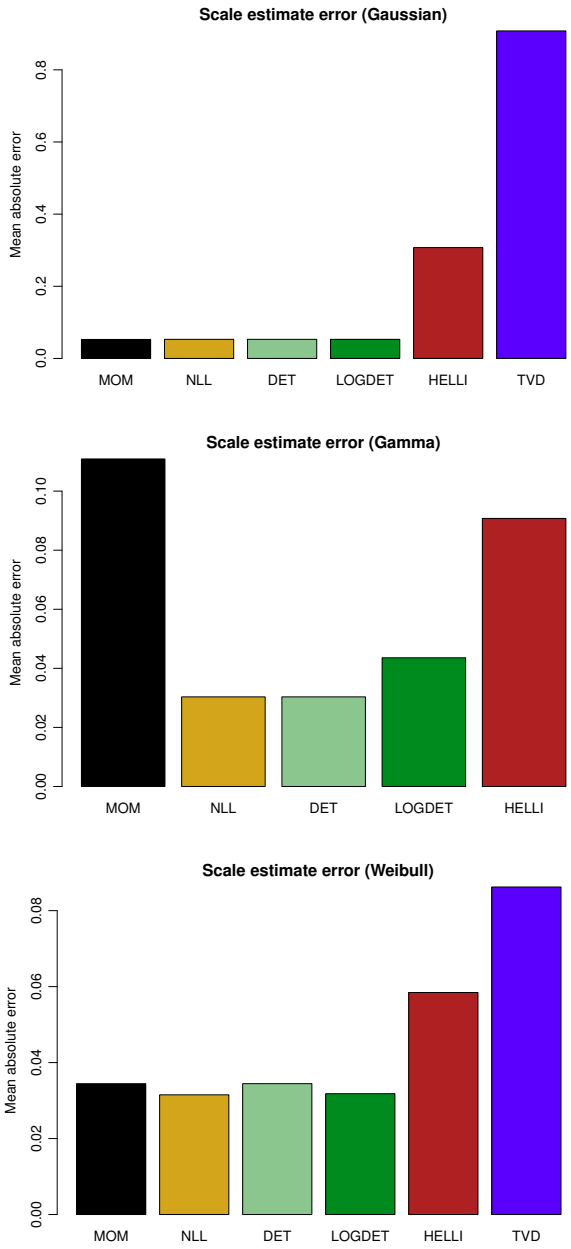

Figure 1: Empirical results of scale estimation for Gaussian, Gamma, and Weibull controlled tests given iid samples and known shift/shape.

methods, as well as evaluation criteria. Section 3.3 gives the empirical results and includes a discussion.

The problem of interest is short-term average wind speed forecasts, where forecasts are (spatially) for the point of measurement. We consider both very short-term forecasts, on the order of seconds, using high-frequency observations, and short-term forecasts on the order of hours. The former is carried out using data from the Heliostat research and development project funded by Google, with a sampling rate of roughly 7.6Hz [21]. The latter is carried out with a large data set assembled from the public AMeDAS observation network data provided by the Japan Meteorological Agency [27]. We have aggregated data from over 1,110 sites

nationwide, and here have filtered this down to 44 sites with exceedingly high-quality data using the following criteria. The minimum average annual wind speed is 2.5m/s, a raw maximum missing value rate is set at 0.1%, and contiguous observations with no missing values or bad data for at least 250 days worth of observations for the year 2012 are required. The sampling period of the raw data is 10 minutes. Geographically, the selected sites span all major inhabited islands of Japan, from Okinawa through to Hokkaido.

The task is formulated as follows. At time t, given a T-length sequence of covariates and scalar responses $\boldsymbol{z}_t \coloneqq ((\boldsymbol{x}_{t-T+1}, y_{t-T+1}), \ldots, (\boldsymbol{x}_t, y_t))$, a target horizon $k > 0$, and a model $\mathcal{P}$, select some $\widehat{P}(k, \boldsymbol{z}_t) \in \mathcal{P}$ to estimate the true conditional $P(y_{t+k}; \boldsymbol{x}_{t+k})$. Point forecasts may be made with $\widehat{y}_{t+k} \coloneqq \text{median}(\widehat{P}(k, \boldsymbol{z}_t))$. The univariate Weibull is an ubiquitous choice for $\mathcal{P}$ in the meteorology literature [44], and we follow this insight here. That is, the model will be fixed, and the comparisons will be across alternative estimation methods. In the prediction tasks for both data sets, horizon will be $k \in \{1, \ldots, 5\}$, though with time-steps being in seconds for the Heliostat case (1s average forecasts), and hours for the AMeDAS case (1h average forecasts).

We use standard references which are very common in the forecasting and time-series literature, additional details are provided in the Appendix. Negative log-likelihood (NLL) [42, 43, 32, 17, 45], continuous ranked probability score (CRPS) [18, 30], and minimum ℓ_m deviation of estimated Weibull CDF from the empirical CDF [29] for $m = 1, 2$ (denoted L1/L2CDF) represent the probabilistic references used. As for deterministic references, random walk, also known as persistence (PER), and an autocorrelation-weighted moving average model from Nielsen et al. [34] are extremely competitive for forecast horizons shorter than a few hours.

Following the simple empirical analysis carried out in Section 3.1, we shall once again make use of the minimal λ_v^α and λ_v^D estimators based on the material in Section 2.2. We shall denote by DET_A, DET_B, and DET_C the λ_v^α-optimizing output for α values of 1.05, 2, and 3. LOGDET denotes the λ_v^D-optimizing output. Weibull shape $\kappa_{t+k} > 0$ is estimated at each t as-is, while scale is modelled auto-regressively as $\xi_{t+k} = \theta_0 + \theta_1 y_t + \cdots + \theta_m y_{t-m+1}$, where $m = k + 3$ for each horizon k. Window length $T > 0$ is set to 3 minutes (Heliostat) and 15 days (AMeDAS). Though other lengths (1–10m, 5–50d) and AR lags 0–5 (in respective time units) were also tested, overall trends in performance metrics remained the same. Thus at forecast horizon $k > 0$, we optimize over $k + 5$ free parameters. We have access to gradients (cf. Section 2.3) and thus the quasi-Newton BFGS method implemented via `optim` in R is used for multivariate optimization.

Four key metrics are used to evaluate density estimation and accuracy of forecasts over the test sets. We use the root mean squared error (RMSE) of signal estimates, long-run volatility difference, R^2 value, and probability of gross error (PGE). If estimate $\widehat{P}(k, \boldsymbol{z}_t)$ yields forecast $\widehat{y}_{t+k}$, then defining $\epsilon_t \coloneqq |y_t - \widehat{y}_t|$, RMSE for a test set of length $N > 0$ is $(\sum_{t=1}^N \epsilon_t^2 / N)^{1/2}$. Long-run volatility difference is simply the absolute value of the difference between the standard deviation (SD) of the observations y_{t+k} and the SD of the forecasts $\widehat{y}_{t+k}$. Let $\widehat{\pi}_{t+k} \coloneqq \widehat{P}(y_{t+k}; k, \boldsymbol{z}_t)$, the cumulative probability assigned by the estimated predictive distribution to the actual observation (recorded k steps later). Let $\overline{P} \coloneqq \sum_{t=1}^N \widehat{\pi}_t / N$. Then R^2 is given as

$$R^2 = \frac{\sum_{t=1}^N (\widehat{\pi}_t - \overline{P})^2}{\sum_{t=1}^N ((\widehat{\pi}_t - \overline{P})^2 + (\pi_t - \overline{P})^2)},$$

where π_t is the empirical cumulative probability of observation y_t. The PGE is defined by $\sum_{t=1}^N I[\epsilon_t \geq \bar{y}]/N$, that is, the relative frequency of exceedingly poor estimates, where $\bar{y}$ is the arithmetic mean (at each given site in AMeDAS case) taken over the test period, and I is the indicator function. Results for the AMeDAS data set discussed in Section 3.3 are averaged over all 44 forecast sites.

3.3 Discussion of empirical results

The main results for both sets of forecasting tasks are given in Fig. 2. We see clearly that all of the proposed methods showed dominant superiority over all probabilistic rivals in terms of RMSE, volatility difference, and PGE, uniformly over all time scales 1–5s and 1–5h. As well, using the most naive possible model, forecast accuracy in terms of RMSE (and similarly MAE, not pictured) for the proposed estimators has already matched or outperformed the strong deterministic references, with no site-specific fine-tuning whatsoever. We note the model fit as gauged by R^2 is better for L1/L2CDF, which should be expected as its parameter optimization effectively maximizes this value. In any case, the model fit found using the proposed methods is far better than when using NLL or CRPS, is maintained on both hour and second time scales, and in addition the volatility difference is by far the smallest, suggesting a more balanced fit up to second-order moments. We note that PGE at the 1–5h scale is within acceptable limits, and gross errors are virtually non-existent at the 1–5s time scale. Probabilistic forecasts at the order of seconds are important for many turbine control applications and yet the literature remains very sparse [28]. The results here may be suggestive of a promising solution to such problems.

Focusing specifically on the AMeDAS network data, in Fig. 3 we have plotted the values of RMSE and R^2 as functions of each site's annual wind speed. The blue-red gradient spans the 1–5 hour forecast horizons, and each row of plots corresponds to a distinct estimation method. An OLS-fit line is also plotted for each horizon length. As a general trend we see that for high velocity sites, forecast error increases and R^2 decreases. One may visually confirm that the rate of performance deterioration of the proposed method is slower than the main probabilistic references, and overall site-to-site volatility is less than NLL and CRPS. To quantify this difference in volatility, we refer to Table 1, which displays the standard deviation of RMSE taken over all the AMeDAS test sites, for each horizon. It is evident that the proposed methods show lower volatility across sites at all horizons, reinforcing the visually confirmed trends of Fig. 3. Very similar trends were found in metrics taken as a function of average annual site temperature, standard deviation of site wind speed, and anemometer height.

Finally, some more general remarks on the proper loss minimizers discussed here. The chief focus of this paper was on a principled approach to constructing proper loss functions for broad classes of parametric models. Note that for many model classes, NLL and CRPS can also be shown to be proper losses, which leads to a natural question: how is one to select between alternative (proper) loss-minimizing estimators? What is the optimal choice given a fixed model and/or a set of observations? As discussed in Section 1, optimality is in general loss-dependent. Even in the well-developed theory of computational learning, where given iid observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N) \sim P$ and some norm-like $\|\cdot\|_A$, and the task of determining an estimate $\widehat{f}_N$ such that some expected loss (risk) $\mathbf{E}_P\|\widehat{f}_N(\boldsymbol{x}) - y\|_A$ is optimized, consistency results (e.g., for empirical error minimizers) in $\|\cdot\|_A$ need not imply any such analogous results for a distinct $\|\cdot\|_B$. In this simple case, we are only interested in predictive capacity, not in a "generative" model descriptive of the underlying phenomena, and thus output evaluation is straightforward. In the more general case, where multiple evaluation metrics (predictive capacity, model descriptiveness, etc.) may be employed, it is clear that a definitive answer to these questions is difficult to expect. The empirical results discussed here are best taken as evidence for the validity of estimators derived using the straightforward approach of Section 2.2, reinforcing their viability as alternatives to more classical quantities. Another interesting question: given a fixed loss used for evaluation after parameters have been estimated, under what conditions, if any, is it unambiguously (or at least with high probability) more desirable to use a different loss for parameter estimation? If such conditions (on the losses and the

Table 1: Sensitivity to spatial condition changes

	SD of RMSE by site				
	1h	*2h*	*3h*	*4h*	*5h*
DET_A	**0.155**	**0.235**	0.289	0.315	0.324
DET_B	0.157	0.237	**0.281**	**0.302**	0.320
DET_C	0.159	0.238	0.284	0.317	0.338
LOGDET	0.175	0.247	0.283	0.303	**0.318**
CRPS	0.280	0.430	0.520	0.580	0.619
NLL	0.502	0.606	0.652	0.681	0.702
L1CDF	0.408	0.408	0.409	0.411	0.412
L2CDF	0.406	0.402	0.401	0.400	0.400
PER	0.159	0.237	0.288	0.330	0.363
NIEL	0.159	0.237	0.288	0.328	0.361

underlying model) exist, can they be approximately checked using data? Much work remains to be done in terms of systematizing the process of loss selection.

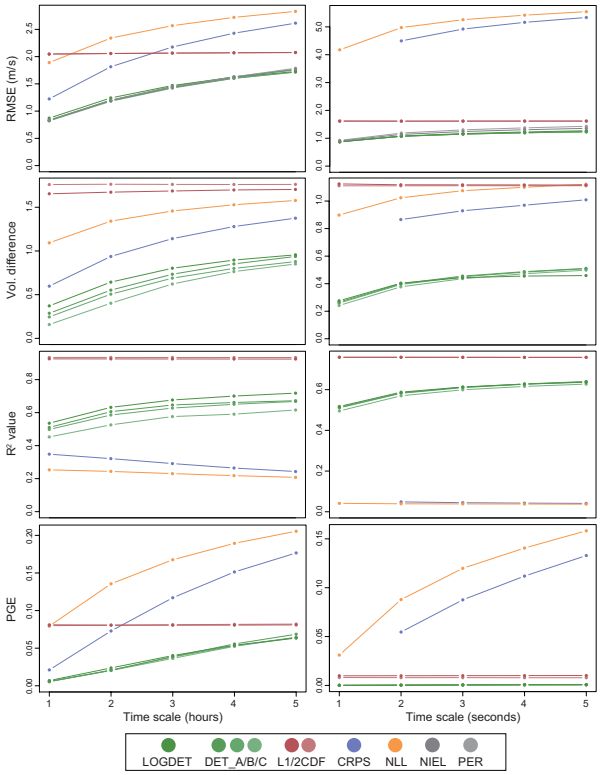

Figure 2: Overall results for all time scales. Left is 1–5h (AMeDAS), right is 1–5s (Heliostat). Various greens are the proposed methods, reds are L1/L2CDF, orange is NLL, blue is CRPS, and greys are PER/NIEL.

4 Conclusion

In this paper, we have presented a systematic new approach to verifying the propriety of losses on parametric models, including sufficient conditions for propriety which require only elementary properties of convex functions. The resulting minimum proper loss estimators can be identified as estimators which minimize a pairwise divergence quantity (between distributions)

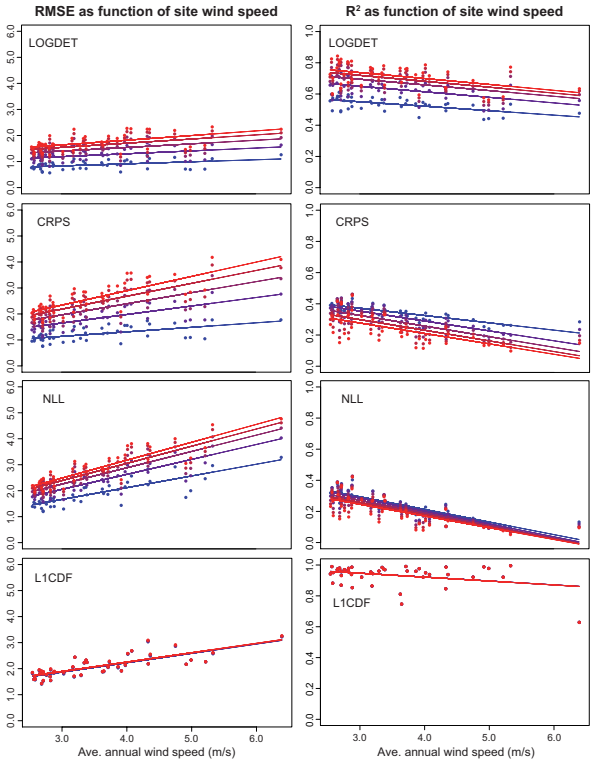

Figure 3: All horizons, RMSE and R^2 as function of wind speed. Blue-red gradient spans 1–5 hour (left) and second (right) time scales.

measured from the underlying distribution. We paid particularly close attention to proper losses requiring only the first and second moments, leading to broad initial propriety results. Illustrative examples, applications to well-known parametric models, an empirical evaluation of simple proper loss classes followed by a discussion of results and a look ahead to future lines of work were provided. While superiority of a particular class of losses in a definite sense is difficult to argue, the estimates based on minimizing the rudimentary, domain-free losses considered here outperformed estimates via numerous standard alternative methods, across several distinct criteria, models, and data sets. Consistency, efficiency, and robustness results for interesting classes of proper loss functions, though inherently more model-specific, are naturally desirable and form an important related task. Similarly, a methodical approach to designing sophisticated losses with the "right" form given the data or a model, perhaps gauged using smoothness and robustness quantifications, also forms another line of work. More generally, as discussed in Section 3.3, the problem of selecting an appropriate loss function from a class of candidates, given observations, remains open; progress on this front can be thought to yield further theoretical grounding to the numerical phenomena observed in this study.

Appendix

Lemma 12. *$\mathbb{S}_+^d$ is an open, convex subset of $\mathbb{R}^{d(d+1)/2}$.*

Proof. Take $\boldsymbol{W}, \boldsymbol{U} \in \mathbb{S}_+^d$. By bilinearity of inner product on real vector spaces, clearly $\alpha\boldsymbol{W} + (1-\alpha)\boldsymbol{U} \in \mathbb{S}_+^d$, so $\mathbb{S}_+^d$ is convex.

With $\boldsymbol{W}$ fixed and $\delta B(\boldsymbol{W})$, the open ball around $\boldsymbol{W}$ with radius $\delta > 0$, note we immediately have $\boldsymbol{U} \in \delta B(\boldsymbol{W}) \implies |w_{i,j} - u_{i,j}| < \delta$ from the norm definition. Let $\boldsymbol{Y} \in \mathbb{S}^d$ be any symmetric

matrix where $|y_{i,j}| \in [0,1]$. Note that $\det(\boldsymbol{W} + \varepsilon \boldsymbol{Y}) = \det \boldsymbol{W} + p(\varepsilon; \boldsymbol{Y})$ where p is a d-degree polynomial in ε. Under the conditions imposed, given ε, p is clearly a bounded function of $\boldsymbol{Y}$. Also noting $\det \boldsymbol{W} > 0$, we may take $\varepsilon > 0$ such that $\det(\boldsymbol{W} + \varepsilon \boldsymbol{Y}) > 0$ for all valid $\boldsymbol{Y}$.

This argument applies identically to the principal minors, allowing that for each principal minor of diagonal length $k = 1, 2 \ldots, d$ we may take a $\varepsilon > 0$ such that the $\det[\boldsymbol{W} + \varepsilon \boldsymbol{Y}]_k > 0$. This implies $\boldsymbol{W} + \varepsilon \boldsymbol{Y} \in \mathbb{S}^d_+$. Fixing ε as the smallest such perturbance taken over $k = 1, \ldots, d$ we note that for all $\boldsymbol{U} \in \varepsilon B(\boldsymbol{W})$, it is clear that there exists some $\boldsymbol{Y}$ as given above such that $\boldsymbol{U} = \boldsymbol{W} + \varepsilon \boldsymbol{Y}$. It follows that $\varepsilon B(\boldsymbol{W}) \subset \mathbb{S}^d_+$, and thus $\mathbb{S}^d_+$ is an open subset of $\mathbb{R}^{d(d+1)/2}$. □

Strict concavity for Corollary 5. Consider strictly concave f on $\mathbb{R}^m$ for integer $m \geq 1$, and let

$$f(\boldsymbol{a}) - f(\boldsymbol{b}) + \langle \boldsymbol{a}^*, \boldsymbol{b} - \boldsymbol{a} \rangle = 0,$$

assuming $\boldsymbol{a} \neq \boldsymbol{b}$. Then using strict concavity we have for $\alpha \in (0,1)$,

$$\begin{aligned} f(\alpha \boldsymbol{a} + (1-\alpha)\boldsymbol{b}) &> \alpha f(\boldsymbol{a}) + (1-\alpha) f(\boldsymbol{b}) \\ &= \alpha f(\boldsymbol{a}) + (1-\alpha)(f(\boldsymbol{a}) + \langle \boldsymbol{a}^*, \boldsymbol{b} - \boldsymbol{a} \rangle) \\ &= f(\boldsymbol{a}) + \langle \boldsymbol{a}^*, (\alpha \boldsymbol{a} + (1-\alpha)\boldsymbol{b}) - \boldsymbol{a} \rangle \\ &\geq f(\alpha \boldsymbol{a} + (1-\alpha)\boldsymbol{b}) \end{aligned}$$

where the last inequality is by the weak supergradient inequality given in the main text, giving us a contradiction, and thus implying $\boldsymbol{a} = \boldsymbol{b}$. In other words, the supergradient inequality holds strictly when concavity holds strictly. □

Strict propriety proof for Example 11. Due to Prop. 9, it is sufficient to prove that $\log\det$ is strictly concave on $\mathbb{S}^d_+$. Fix some $\boldsymbol{W} \in \mathbb{S}^d_+$, and let $\boldsymbol{Y}$ be any symmetric matrix with elements $|y_{i,j}| \in [-1,1]$. Let $g(u; \boldsymbol{Y}) \coloneqq \log\det(\boldsymbol{W} + u\boldsymbol{Y})$. As noted in Lemma 12, by openness of $\mathbb{S}^d_+$, there exists $\varepsilon > 0$ such that $\varepsilon B(\boldsymbol{W}) \subset \mathbb{S}^d_+$. If $u \in (-\varepsilon/d, \varepsilon/d)$, then clearly $\boldsymbol{W} + u\boldsymbol{Y} \in \varepsilon B(\boldsymbol{W})$ and is thus positive definite. We thus set $\mathrm{dom}(g) = (-\varepsilon/d, \varepsilon/d)$, from which it follows that $g(u; \boldsymbol{Y})$ is differentiable on $\mathrm{dom}(g)$.

Let $\sqrt{\boldsymbol{W}} = \boldsymbol{S}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{S}^T$, where $\boldsymbol{W} = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^T$ is the spectral decomposition of $\boldsymbol{W}$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\eta_1, \ldots, \eta_d)$. It is thus immediate that $\sqrt{\boldsymbol{W}} \in \mathbb{S}^d_+$ as eigenvalues are the $\sqrt{\eta_i} > 0$. Let $\boldsymbol{\Gamma} \coloneqq \sqrt{\boldsymbol{W}}^{-1} \boldsymbol{Y} \sqrt{\boldsymbol{W}}^{-1} \in \mathbb{S}^d$, with eigenvalues γ_i. Then we have

$$\begin{aligned} g(u; \boldsymbol{Y}) &= \log\det \sqrt{\boldsymbol{W}}(\boldsymbol{I} + u\boldsymbol{\Gamma})\sqrt{\boldsymbol{W}} \\ &= \log\det \boldsymbol{W} + \sum_{i=1}^{d} \log(1 + u\gamma_i). \end{aligned}$$

As second derivative is $\partial^2 g(u; \boldsymbol{Y}) = -\sum \gamma_i^2/(1+u\gamma_i)^2 < 0$, we have that $g(u; \boldsymbol{Y})$ is strictly concave on its domain [41, Thm. 4.4]. If denote $\boldsymbol{B}_u \coloneqq \boldsymbol{W} + u\boldsymbol{Y}$, then by definition of g can readily confirm

$$\log\det(\alpha \boldsymbol{B}_t + (1-\alpha)\boldsymbol{B}_u) > \alpha \log\det \boldsymbol{B}_t + (1-\alpha)\log\det \boldsymbol{B}_u$$

for any $t \neq u$ on $\mathrm{dom}(g)$, and any $\boldsymbol{Y}$ as described. We now let $\zeta \coloneqq \varepsilon/2d$, so $[-\zeta, \zeta] \subset \mathrm{dom(g)}$, and clearly for any $\boldsymbol{B} \in \zeta B(\boldsymbol{W})$ there exists $u \in \mathrm{dom}(g)$ and $\boldsymbol{Y}$ such that $\boldsymbol{B} = \boldsymbol{W} + u\boldsymbol{Y}$. The strict concavity inequality above thus implies $\log\det$ is concave over $\zeta B(\boldsymbol{W}) \subset \varepsilon B(\boldsymbol{W})$, and as $\boldsymbol{W}$ was arbitrary, the result follows. □

Reference estimators

Given the data and task in Section 3, we seek standard estimators used in the domain considered which function as reasonable benchmarks. Most ubiquitous in the literature is the negative log-likelihood (NLL) estimator

$$\widehat{\boldsymbol{\theta}}_{NLL} \coloneqq \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} (-1) \log p(\boldsymbol{x}_t; \boldsymbol{\theta})$$

used in wind speed applications finding both an optimal fit [42, 32, 45] and tasks focused solely on forecasting [43, 17]. Alternatively, the continuous ranked probability score (CRPS) defined for CDF P by

$$\mathrm{CRPS}(P(\boldsymbol{\theta}), x) \coloneqq \int |P(u; \boldsymbol{\theta}) - I[u \geq x]|^2 \, du$$

has been used to define estimator $\widehat{\boldsymbol{\theta}}_{CRP} \coloneqq \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \mathrm{CRPS}(P(\boldsymbol{\theta}), x_t)$ over a sliding window of size $T > 0$, reported to be a successful and robust replacement for NLL in the forecasting literature recently [18, 30]. Gaussian models readily admit closed-form expressions for the CRPS integral [20], and in the case of a Weibull model we may arrive at such a form by detouring through a particular generalized extreme value distribution [24]. We note both CRPS and NLL are proper loss functions [19]. The standard deterministic reference is the random walk, also known as persistence (PER), defined $\widehat{x}_{t+k} = x_t$, and is traditionally extremely competitive for short-term forecasts. Another very strong deterministic reference comes from Nielsen et al. [34], here denoted NIEL, and is a simple autocorrelation-weighted moving average model formulated as $\widehat{x}_{t+k} = \rho_k x_t + (1-\rho)\bar{x}$, where ρ_k is the k-lagged autocorrelation coefficient, and $\bar{x}$ is the arithmetic mean of recent observations. For a classical probabilistic reference dating back at least as far as Justus et al. [29], taking two consecutive logarithms of the Weibull CDF $W(x; \xi, \kappa)$ yields $\log(-\log(1-\pi_t)) = \kappa \log x_t + \kappa \log \xi$, where π_t denotes the empirical cumulative probability assigned to observation x_t. Let $\boldsymbol{w} = (w, \boldsymbol{\omega})$, and let us model $\kappa = w$ as a scalar and $\xi = \exp(\boldsymbol{\omega}^T \boldsymbol{\phi}_t)$, where $\boldsymbol{\phi}_t = \boldsymbol{\phi}(\boldsymbol{X}_t)$ is a given vector of covariates. If using the ℓ_m norm on $\mathbb{R}$ we define $(\widehat{w}_m, \widehat{\boldsymbol{\omega}}_m) \coloneqq \arg\min_{\boldsymbol{w}} \|\tilde{\boldsymbol{\pi}} - \boldsymbol{\Phi}\boldsymbol{w}\|_m$, then the L1CDF and L2CDF estimators are defined by $\widehat{\boldsymbol{\theta}}_m \coloneqq (\widehat{\boldsymbol{\omega}}_m / \widehat{w}_m, \widehat{w}_m)$ for $m = 1, 2$. Naturally $m = 2$ is standard OLS estimation, while $m = 1$ is minimum mean absolute deviation estimation, and efficient standard algorithms for this computation are due to Barrodale and Roberts [5].

Experimental data

We have included the raw data used in the performance evaluation section of this paper. All data are compressed into R data files (extension `.rda`) and may be readily accessed using the language and environment R, freely available for all major operating systems at `http://www.r-project.org/`. If we assume the files are saved in `tmp` in the user's home directory, with R installed and running, they may be loaded in R by

```
~/ $ R
...
> setwd(dir="~/tmp")
> load("amedas_data.rda")
> load("heliostat_data.rda")
```

The data are stored in dictionary-like R `list` structures. For example, the observations for the ith (filtered) AMeDAS site and the jth observation day of the Heliostat data can be specified using

```
> i_data_am <- amedas_data[[i]]
> j_data_hs <- heliostat_data[[j]]
```

where we assume `i` and `j` have been assigned some positive integer. The AMeDAS observations are unprocessed 10min observations (free of missing/bad values as discussed in the paper), while the Heliostat data provided is the pre-smoothed set of observations, where disjoint seven time-step subsets are taken and averaged in sequential order, thereby resulting in a set of per-second average velocities. Only days with full-day observations were considered, and thus the first and last observation days were excluded from the attached dataset.

We have also included the filtered AMeDAS site indices in the file called

```
filtered\_sites.rda
```

which take the form of a two-column matrix of character strings (two concatenated arrays with a built-in two-dimensional index):

```
> filtered_sites
      AMEMGR_DATA_PREF AMEMGR_DATA_BLOCK
 [1,] "11"             "1054"
 [2,] "13"             "0028"
 ...
 [43,] "91"             "1152"
 [44,] "91"             "1354"
```

where the first column contains prefecture codes, and the second contains site-identifying "block" codes. It is possible for two distinct sites in different prefectures to have the same block code, though we note that every site is uniquely specified by this pair of numbers. The official AMeDAS site names attached to each two-value ID corresponding to the filtered sites are as follows:

```
"11_1054" "Toyotomi"       "55_0546" "Tomari"
"13_0028" "Mashike"        "56_0564" "Shika"
"14_1085" "Ishikari"       "57_1316" "Koshino"
"14_1193" "Hamamasu"       "63_0625" "Akashi"
"14_1459" "Chitose"        "63_1337" "Nandan"
"14_1507" "Ebetsu"         "67_0686" "Takehara"
"17_0065" "Yubetsu"        "71_1242" "Gamoda"
"17_0076" "Shari"          "81_0775" "Hofu"
"24_1199" "Okushiri"       "81_0939" "Yuya"
"43_1070" "Tokorozawa"     "82_1141" "Dazaifu"
"44_0366" "Hachioji"       "83_0796" "Musashi"
"44_0370" "Edogawa-Rinkai" "84_1138" "Arikawa"
"44_0371" "Haneda"         "84_1144" "Ashibe"
"45_0384" "Kamogawa"       "87_1481" "Akae"
"48_0415" "Nobeyama"       "88_0898" "Kaminaka"
"51_0470" "Toyohashi"      "88_0899" "Onoaida"
"51_0984" "Minami-Chita"   "88_1540" "Nakanoshima"
"52_0483" "Hagiwara"       "91_0901" "Oku"
"54_0521" "Niitsu"         "91_0909" "Itokazu"
"54_0522" "Maki"           "91_1145" "Ibaruma"
"54_0532" "Kashiwazaki"    "91_1152" "Tokashiki"
"54_1315" "Hajikizaki"     "91_1354" "Hateruma"
```

Other related datasets for these and all other sites on the AMeDAS network are made publicly available by the Japan Meteorological Agency at `http://www.jma.go.jp/`.

References

[1] Anand, C. S. and Sahambi, J. S. (2010). Wavelet domain non-linear filtering for MRI denoising. *Magnetic Resonance Imaging*, 28(6):842–861.

[2] Artin, E. (1964). *The Gamma Function*. Holt, Rinehart and Winston.

[3] Ash, R. B. and Doleans-Dade, C. (2000). *Probability and measure theory*. Academic Press.

[4] Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

[5] Barrodale, I. and Roberts, F. D. K. (1973). An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5):839–848.

[6] Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

[7] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

[8] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150.

[9] Broniatowski, M. and Keziou, A. (2009). Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36.

[10] Csiszár, I. (1972). A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1):191–213.

[11] Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273.

[12] Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93.

[13] Dawid, A. P., Lauritzen, S., and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608.

[14] Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *The Annals of Statistics*, pages 65–81.

[15] Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, 11(3):793–803.

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

[17] Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7):579–594.

[18] Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space-time method. *Journal of the American Statistical Association*, 101(475):968–979.

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

[20] Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

[21] Google (2014). *REC Initiative*.

[22] Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, pages 1367–1433.

[23] Halmos, P. R. (1974). *Measure Theory, Graduate Texts in Mathematics (18)*. Springer-Verlag New York.

[24] Holland, M. J. and Ikeda, K. (2014). Forecasting in wind energy applications with site-adaptive Weibull estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, Florence, Italy.

[25] Holland, M. J. and Ikeda, K. (to appear). Location robust estimation of predictive Weibull parameters in short-term wind speed forecasting. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, Brisbane, Australia.

[26] Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

[27] Japan Meteorological Agency (2013). *About AMeDAS*.

[28] Jiang, Y., Song, Z., and Kusiak, A. (2013). Very short-term wind speed forecasting with Bayesian structural break model. *Renewable Energy*, 50:637–647.

[29] Justus, C. G., Hargraves, W. R., and Yalcin, A. (1976). Nationwide assessment of potential output from wind-powered generators. *Journal of Applied Meteorology*, 15(7):673–678.

[30] Lerch, S. and Thorarinsdottir, T. L. (2013). Comparison of nonhomogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, 65.

[31] Matsubara, T., Gómez, V., and Kappen, H. J. (2014). Latent Kullback Leibler control for continuous-state systems using probabilistic graphical models. In *30th Conference on Uncertainty in Artificial Intelligence (UAI-2014)*.

[32] Morgan, E. C., Lackner, M., Vogel, R. M., and Baise, L. G. (2011). Probability distributions for offshore wind speeds. *Energy Conversion and Management*, 52(1):15–26.

[33] Newby, M. J. (1980). The properties of moment estimators for the weibull distribution based on the sample coeffkient of variation. *Technometrics*, 22(2):187–194.

[34] Nielsen, T. S., Joensen, A., Madsen, H., Landberg, L., and Giebel, G. (1998). A new reference for wind power forecasting. *Wind Energy*, 1(1):29–34.

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

[36] Permuter, H., Francos, J., and Jermyn, H. (2003). Gaussian mixture models of texture and colour for image database retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003)*, volume 3, pages 569–572.

[37] Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, 14(1):249–272.

[38] Pham, D. T. and Garat, P. (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725.

[39] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[40] Rinne, H. (2010). *The Weibull distribution: A handbook*. CRC Press.

[41] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

[42] Seguro, J. V. and Lambert, T. W. (2000). Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis. *Journal of Wind Engineering and Industrial Aerodynamics*, 85(1):75–84.

[43] Taylor, J. W., McSharry, P. E., and Buizza, R. (2009). Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, 24(3):775–782.

[44] 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: Series A (Statistics in Society)*, 173(2):371–388.

[45] Zhang, J., Chowdhury, S., Messac, A., and Castillo, L. (2013). A multivariate and multimodal wind distribution model. *Renewable Energy*, 51:436–447.