TI 2026-022/III
Tinbergen Institute Discussion Paper

Proper and Robust Autoregressive Derivative Adaptive Models

Ramon de Punder[1]

1 University of Amsterdam and Tinbergen Institute

Proper and Robust Autoregressive Derivative Adaptive Models*

Ramon de Punder†
Department of Quantitative Economics
University of Amsterdam and Tinbergen Institute

May 15, 2026

Abstract

This paper introduces the class of Proper and Robust Autoregressive Derivative Adaptive (PRADA) models, extending score-driven updates beyond the logarithmic scoring rule to all strictly proper and locally proper scoring rules and strictly consistent scoring functions. PRADA updates reduce an expected local divergence measure under misspecification and thereby generalize the information-theoretic foundation of score-driven models beyond the Kullback-Leibler divergence. They are interpreted as the online analogues of M-estimators, and are linked to online Z-estimation through strict identification functions. When derived from scoring functions or identification functions, PRADA updates operate directly on elicitable functionals of the postulated conditional distribution, such as conditional means, quantiles or risk measures, and therefore do not require a parametric model. The results provide general conditions under which updates are guaranteed to reduce their corresponding divergence, establish robustness through bounded and censored updates, and encompass many existing score-driven inspired models as special cases.

Keywords: Generalized autoregressive score (GAS); Dynamic conditional score (DCS); Scoring rules; Scoring functions; Divergence measures; Censoring

*The author gratefully acknowledges Mathijs Dijkstra, Cees Diks, Timo Dimitriadis, Peter Grünwald, Andrew Harvey, Yi He, Tjeert Keijzer, Frank Kleibergen, Roger Laeven, André Lucas, Rutger-Jan Lange, Dario Palumbo, Johannes Resin, Dick van Dijk and Chen Zhou for their comments and suggestions.

†Corresponding author. Mailing Address: PO Box 15867, 1001 NJ Amsterdam, The Netherlands. Phone: +31 (0) 20 525 4252. Email: r.f.a.depunder@uva.nl.

1 Introduction

In time series modeling, the introduction of score-driven updates by Harvey (2013) and Creal et al. (2013) has spurred a growing interest in their applications. This, in turn, has led to an expanded analysis of the associated statistical properties of this model class, including the analysis in Gorgi et al. (2024) and De Punder et al. (2026), as well as to numerous extensions of the original framework (Patton et al. 2019; Harvey and Ito 2020; Blasques et al. 2023; Creal et al. 2024). However, while stationarity, ergodicity and statistical inference have been established for broad classes of updating rules (Gorgi 2020; Blasques et al. 2023), the information-theoretic foundation developed in De Punder et al. (2026) remains confined to the log-likelihood framework. Moreover, since loss functions must preserve a meaningful link with the underlying distribution (Gneiting and Raftery 2007; Gneiting 2011), not all parameter updates can rely on such a foundation.

In this paper, we extend the class of score-driven updates to the broader class of so-called Proper and Robust Autoregressive Derivative Adaptive (PRADA) updates. PRADA updates allow for replacing the derivative of the logarithmic scoring rule by the derivative of any other loss function that preserves the theoretical foundation of score-driven updates under an alternative divergence measure. More specifically, whereas score-driven updates and their equivalents are unique in reducing the Expected Kullback and Leibler (1951) divergence between the postulated time-varying density and the true density, PRADA updates are unique in reducing some fixed expected (local) divergence measure. This encompasses updates derived from strictly proper scoring rules, strictly consistent scoring functions, and strictly locally proper weighted scoring rules. Many existing updates inspired by the score-driven framework are PRADA updates for which this paper provides the theoretical foundation, while also motivating the design of new, theoretically grounded

updating rules.

A key motivation for departing from the logarithmic scoring rule is its sensitivity to outliers (Selten 1998), as it penalizes events that are extremely unlikely under correct specification with losses tending to infinity. While such confidence in the postulated model can be desirable for model discrimination (Neyman and Pearson 1933), robustness to model misspecification often proves valuable in practical applications. Donker van Heel et al. (2025) show that the sensitivity of the logarithmic scoring rule carries over to the instability of the score-driven filter, resulting in mean-squared-error-unbounded updates in commonly-used settings such as Gaussian volatility models and other non-Lipschitz continuous scores. Similarly, De Punder et al. (2026) introduce clipped variants of score-driven updates for cases in which global boundedness of the expected Hessian is not guaranteed under the assumed model. In this paper, we additionally show that replacing the score with a quasi-score, as per Blasques et al. (2023), does not necessarily yield a reduction in the expected Kullback-Leibler (EKL) divergence. A more direct approach is to retain the model specification and instead replace the logarithmic scoring rule with other proper but more robust alternatives such as the Continuous Ranked Probability Score (CRPS) of Matheson and Winkler (1976) and the censored likelihood score (CSL) proposed by Diks et al. (2011). At this point, we address the terminological ambiguity that the term "score" refers both to the outcome of a scoring rule *and* the gradient of the logarithmic scoring rule. The intended meaning, however, is always clear from the context.

In a Gaussian location model the CRPS corresponds to a smooth version of the Huber estimator (Huber 1981, p. 208), used to robustly estimate the population mean (Gneiting and Raftery 2007). Estimating parameters based on the CRPS serves as a robust alternative to traditional maximum-likelihood estimation and fits the framework of M-

estimation (Huber 1964). An important difference with M-estimation is the online target, custom in time-series applications. Viewing score-driven updates as the online version of maximum-likelihood, the consistency of which relies on the strict propriety of the logarithmic scoring rule (Wald 1949; Gneiting and Raftery 2007), PRADA updates can be seen as the online version of M-estimation under similar restrictions as required for consistency of M-estimators.

To accommodate weighted scoring rules and scoring functions, we consider a broader class of loss functions beyond that of the strictly proper scoring rules of Gneiting and Raftery (2007). We show that the generalized class of strictly locally proper scoring rules of Holzmann and Klar (2017) can be formulated in terms of strict propriety with respect to some functional of the distribution, parameterized by a weight function on the outcome space. Following the decision-theoretic perspectives of Grünwald and Dawid (2004), Dawid (2007), Gneiting (2011) and Dawid and Musio (2014), we establish a unifying concept for loss functions defined on different action domains, e.g., scoring functions on outcome spaces or scoring rules on classes of distributions, which we term strict local propriety with respect to a functional, generalizing Theorem 3 in Gneiting and Ranjan (2011). We correspondingly extend the notion of local divergences with respect to a weight function, as in Definition 2 of De Punder et al. (2025), to local divergences with respect to a functional, identifying the true distribution up to the chosen functional.

In less general contexts, the importance local divergences introduced by De Punder et al. (2025) builds on a broad statistical literature (Gneiting and Raftery 2007; Diks et al. 2011; Gneiting and Ranjan 2011; Gneiting 2011; Holzmann and Klar 2017). The adverse consequences of parameter updates that do not reduce a local divergence mirror those of forecast evaluations based on improper scoring rules such as the linear or weighted likelihood

scoring rule of Amisano and Giacomini (2007). As illustrated by Example 2 in Section 2, parameter updates based on the linear scoring rule can become entirely independent of the true distribution. This underscores that the general updating rules of Gorgi (2020) and Blasques et al. (2023) require additional guidance; only those intersecting with the PRADA class are guaranteed to reduce a local divergence with respect to the true distribution.

Many PRADA models were ahead of their theoretical formalization. Catania et al. (2025) consider updates based on the CRPS to achieve more robust updating rules. Score-driven models for count data and circular data, as discussed by Gorgi (2020) and Harvey et al. (2024), respectively, are naturally encompassed by the generality of the outcome space. Other types of updates also arise. The Hyvärinen (2005) scoring rule is second-order local, as it depends on the likelihood through its first and second derivatives only, and is therefore conceptually related to score-driven updates. The PseudoSpherical and Power scoring rule families include the logarithmic scoring rule as limiting cases when their tuning parameter tends to one, while providing more robust updating behavior for other parameter values. Furthermore, the CRPS belongs to the broader class of kernel scores, which accommodate highly general outcome spaces, and unlike likelihood-based updates can be applied to distributions characterized only through their distribution function or characteristic functions; see, for instance Example 12 in Gneiting and Raftery (2007).

Harvey and Ito (2020) introduce an example of a PRADA update based on the CSL, guaranteed to reduce the expected censored Kullback-Leibler divergence. De Punder et al. (2025) demonstrate that censoring can be applied to any other strictly proper scoring rule while preserving the corresponding unweighted divergence measure. Weighting strictly proper scoring rules, including those mentioned above, can be useful for downweighting extreme observations, thereby enhancing the robustness of the updating step and simul-

taneously addressing the issue of missing values (Harvey and Ito 2020). Our extension to strict local propriety with respect to functionals further accommodates weighted scoring rules beyond the class of strictly locally proper scoring rules, including updates based on the threshold-weighted CRPS of Gneiting and Ranjan (2011) for center indicator functions, and, analogously, the dynamic Tobit model proposed by Harvey and Liao (2023).

In the context of strictly consistent scoring functions, score-driven-inspired updates include the updating rule for the elicitable (Value-at-Risk, Expected Shortfall) pair proposed by Patton et al. (2019), as well as, the updating rule for the vector of quantiles introduced by Catania and Luati (2023). By Osband's principle (Fissler and Ziegel 2016), PRADA updates can often be expressed in terms of identification functions, solutions to which we interpret as an online Z-estimation problem. This formulation allows for direct incorporation of identification functions that are often readily available (Gneiting and Ranjan 2011; Dimitriadis et al. 2024).

We proceed as follows. Section 2 introduces preliminaries on updating rules, loss functions and local divergences. Section 3 presents the main theoretical results, with examples for unweighted and weighted scoring rules included in Sections 3.4 and 3.5, respectively. Section 4 examines the implications for strictly consistent scoring functions and strict identification functions. Section 5 concludes. Code for the reproduction of included figures is available from the paper's GitHub: `https://github.com/rdpunder/PRADA`.

2 Updating rules and divergence measures

Consider a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a standard Borel measurable space $(\mathcal{Y}, \mathcal{G})$. For $t = 1, \ldots, T$, with $T \in \mathbb{N}$, define random variables $Y_t : (\Omega, \mathcal{F}) \to (\mathcal{Y}, \mathcal{G})$ and write $Y_{1:T} := (Y_1, \ldots, Y_T) : \Omega \to \mathcal{Y}^T$ with product σ-algebra $\mathcal{G}^{\otimes T}$. The natural filtration is

$\{\mathcal{F}_t\}_{t=1}^T$ with $\mathcal{F}_t := \sigma(Y_1, \ldots, Y_t)$. We denote the true conditional distribution of Y_t given $\mathcal{F}_{t-1}$ by the kernel $\mathrm{P}_t(\cdot) := \mathbb{P}\big(Y_t \in \cdot \,\big|\, \mathcal{F}_{t-1}\big)$ on $(\mathcal{Y}, \mathcal{G})$, with distribution function P_t and, whenever it exists, μ_t-density $p_t(y) = \frac{\mathrm{dP}_t}{\mathrm{d}\mu_t}(y)$. Here, we are explicit about the underlying measurable space to allow for outcome spaces beyond Euclidean space and distributions that are not Lebesgue dominated, such as the continuous-discrete censored distribution introduced by De Punder et al. (2025).

The conditional distribution implied by the postulated model is denoted by $\mathrm{F}_{t|t-1}$, driven by a time-varying parameter $\vartheta_{t|t-1} \in \Theta \subseteq \mathbb{R}^k$. As in De Punder et al. (2026), we assume that Θ is open and convex. An *updating rule* $\phi : \mathcal{Y} \to \Theta$ determines how parameters $\vartheta_{t|t-1}$ are updated to $\vartheta_{t|t} := \phi(y_t, \vartheta_{t|t-1})$ based on a realization y_t, and induces the corresponding update of the conditional distribution from $\mathrm{F}_{t|t-1} \equiv \mathrm{F}_{\vartheta_{t|t-1}}$ to $\mathrm{F}_{t|t} \equiv \mathrm{F}_{\vartheta_{t|t}}$. Updating rules under consideration therefore yield *observation-driven* models (Cox 1981). The conditional distribution is allowed to depend on static parameters or exogenous variables as well, though suppressed in the notation. Its distribution and, whenever it exists, μ_t-density, are denoted by $F_{t|t-1} \equiv F_{\vartheta_{t|t-1}}$ and $f_{t|t-1} \equiv f(\cdot|\vartheta_{t|t-1})$, respectively. The time-varying parameter $\vartheta_{t|t-1}$ typically represents either a parameter vector indexing a parametric family or a functional of the distribution. Whenever flexibility of the parameter space is desirable, dependence on $\vartheta_{t|t-1}$ can be represented through a *link function* $\Theta \mapsto \widetilde{\Theta}$, defined as any invertible and continuous mapping, with $\widetilde{\Theta} \subseteq \mathbb{R}^k$ (Creal et al. 2013, Sec. 2.4). Following De Punder et al. (2025), we assume that the support of $\mathrm{F}_{t|t-1}$ contains the support of P_t.

Whenever two random variables, densities or distribution functions are said to be equal, this is understood to hold almost surely. Moreover, two measures $\mathrm{F}_{t|t-1}$ and $\mathrm{G}_{t|t-1}$ coinciding on $\mathcal{G}$, that is, $\mathrm{F}_{t|t-1}(E) = \mathrm{G}_{t|t-1}(E), \forall E \in \mathcal{G}$, is denoted by $\mathrm{F}_{t|t-1} = \mathrm{G}_{t|t-1}$. Two measures coinciding on $(R \subseteq \mathcal{Y}) \in \mathcal{G}$, that is, $\mathrm{F}_{t|t-1}(E \cap R) = \mathrm{G}_{t|t-1}(E \cap R)$, $\forall E \in \mathcal{G}$, is

similarly denoted $\mathrm{F}_{t|t-1} \overset{R}{=} \mathrm{G}_{t|t-1}$. For a matrix $A \in \mathbb{R}^{k\times k}$, we denote the trace by $\mathrm{tr}(A)$, the operator norm by $\|A\|_2$, which is induced by the Euclidean norm $\|x\|_2$ for vectors $x \in \mathbb{R}^k$, and positive (semi)definiteness by $A \succ (\succeq)\ O_k$, where O_k is the $k \times k$ zero matrix and $A \in \mathbb{R}^{k\times k}$. Negative (semi)definiteness is analogously denoted $A \prec (\preceq)\ O_k$. The $k \times k$ identity matrix is denoted I_k. For a function $\bar{g} : \mathbb{R}^l \to \mathbb{R}$, with $l \in \mathbb{N}$, we denote the gradient, Hessian and Laplacian by $\nabla_x \bar{g}(x) := \frac{\partial}{\partial x}\bar{g}(x)$, $\nabla_x^2 \bar{g}(x) := \frac{\partial^2}{\partial x \partial x^\top}\bar{g}(x)$ and $\Delta_x \bar{g} := \mathrm{tr}\big(\nabla_x^2 \bar{g}(x)\big)$, respectively. Moreover, the gradient of a vector-valued function $\bar{g}_m : \mathbb{R}^l \to \mathbb{R}^m$, with $m > 1$, is the $m \times l$ Jacobian matrix $\nabla_x \bar{g}(x)$ with elements $[\nabla_x \bar{g}(x)]_{i,j} = \frac{\partial g_i(x)}{\partial x_j}$. We suppress time dependence in ∇_x, which will be clear from the context.

2.1 Local divergence space

PRADA updates will be defined with respect to a *local divergence space* $(\mathcal{P}, \mathbb{D}, \mathcal{T})$, where $\mathbb{D}$ denotes a *local divergence* with respect to a *functional* $\mathcal{T}$. We carefully develop the concept of a local divergence space. For notational convenience, we temporarily suppress time indices.

Adopting the decision-theoretic perspective of Grünwald and Dawid (2004), Dawid (2007), Gneiting (2011) and Dawid and Musio (2014), we define a *loss function* $L : \mathcal{A} \times \mathcal{Y} \to \bar{\mathbb{R}}$ for actions $a \in \mathcal{A}$ and outcomes $y \in \mathcal{Y}$. Throughout, we assume that losses are negatively oriented, F-integrable for all $\mathrm{F} \in \mathcal{P}$, and $\mathcal{G}$-measurable. We consider two types of losses: *scoring rules* $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, and *scoring functions* $\ell : \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$, where $\mathcal{X} \subseteq \mathbb{R}^{\tilde{k}}$, with $\tilde{k} \in \mathbb{N}$. In the context of scoring functions, we additionally require $\mathcal{Y} \subseteq \mathbb{R}^l$.

An action $a_{\mathrm{P}}^* \in \mathcal{A}$ is called a *Bayes act* under $\mathrm{P} \in \mathcal{P}$ if it is a minimizer of the P-expected loss, that is, $a_{\mathrm{P}}^* \in \arg\min_{a\in\mathcal{A}} \mathbb{E}_{\mathrm{P}} L_{\mathcal{A}}(a, Y)$. A scoring rule is termed *proper* with respect to $\mathcal{P}$ if the true distribution P is a Bayes act, for all $\mathrm{P} \in \mathcal{P}$, and *strictly proper* with

respect to $\mathcal{P}$ if it is the unique Bayes act (up to P-a.s. equivalence), for all $\mathrm{P} \in \mathcal{P}$. Similarly, for a potentially set-valued functional $\tilde{u} : \mathcal{P} \twoheadrightarrow \mathcal{Y}$, a scoring function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is called *consistent* with respect to $(\tilde{u}, \mathcal{P})$ if every $a_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$ is a Bayes act, for all $\mathrm{P} \in \mathcal{P}$, and *strictly consistent* with respect to $(\tilde{u}, \mathcal{P})$ if only every $a_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$ is a Bayes act, for all $\mathrm{P} \in \mathcal{P}$ (Gneiting 2011; Fissler and Ziegel 2016). A functional $\tilde{u}$ that admits a strictly consistent scoring function ℓ is called *elicitable*.

In applications, single-valued functionals $u : \mathcal{P} \to \mathcal{Y}$ are more common. We therefore focus on this case and defer the treatment of set-valued functionals to Appendix D. For instance, when a quantile $q_\beta(\mathrm{F}) := \{\lim_{t \uparrow x} F(t) \leq \beta \leq F(x)\}$, with $\beta \in (0, 1)$, is set-valued, one may instead work with its associated Value-at-Risk, $\mathrm{VaR}_\beta(\mathrm{F}) := \inf q_\beta(\mathrm{F})$.

We also consider functionals of the form $v : \mathcal{P} \to \mathcal{Q}$, where $\mathcal{Q}$ denotes a class of distributions on $(\mathcal{Y}, \mathcal{G})$. A prominent example is the censoring functional $v_R^\flat(\cdot)$ introduced by De Punder et al. (2025), which maps $\mathrm{F} \in \mathcal{P}$ to censored distributions

$$\mathrm{F}_R^\flat(E) \equiv v_R^\flat \# \mathrm{F}(E) := \mathrm{F}(E \cap R) + \mathrm{F}(R^c)\delta_*(E), \qquad \forall E \in \mathcal{G}^\flat,$$

where $\mathcal{G}^\flat := \sigma(\{\mathcal{G}, *\})$, $* \notin \mathcal{Y}$ is an abstract censoring point extending the outcome space, and $(R \subseteq \mathcal{Y}) \in \mathcal{G}$. When discussing properties common to both functional types, we refer to the class of functionals encompassing u and v as $\mathcal{T} : \mathcal{P} \to \mathcal{V}$, with $\mathcal{V} \in \{\mathcal{Y}, \mathcal{Q}\}$.

The censoring functional can be used to reformulate the notion of *strict local propriety* introduced by Holzmann and Klar (2017), who discuss scoring rules localized by some measurable weight function $w : \mathcal{Y} \to [0, 1]$, to emphasize *regions of interest* $R_w := \{y \in \mathcal{Y} : w(y) > 0\}$. They define a weighted scoring rule $S_w : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ to be *localizing* if $\mathrm{F} \overset{R_w}{=} \mathrm{G}$ implies $S(\mathrm{F}, y) = S(\mathrm{G}, y)$, for all $y \in \mathcal{Y}$ and generalize the notion of strict propriety to *strict local propriety* with respect to $(\mathcal{P}, w)$. The latter is equivalent to S_w being strictly locally proper with respect to $(\mathcal{P}, v_{R_w}^\flat)$.

Definition 1. *A scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is strictly locally proper with respect to $(\mathcal{P}, \mathcal{T})$ if $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) := \mathbb{E}_\mathrm{P} S(\mathrm{F}, Y) - \mathbb{E}_\mathrm{P} S(\mathrm{P}, Y) \geq 0$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$, with strict equality if and only if $\mathcal{T}(\mathrm{P}) = \mathcal{T}(\mathrm{F})$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$.*

Defining $S_\ell(\mathrm{F}, y) := \ell(u(\mathrm{F}), y)$, it follows that S_ℓ is strictly locally proper with respect to $(\mathcal{P}, u)$ if ℓ is strictly consistent with respect to $(\mathcal{P}, u)$, generalizing Theorem 3 in Gneiting and Ranjan (2011) for point-valued functionals. Section 4 elaborates on this special case, which we illustrate with the following example.

Example 1. *Let $\mathcal{P}_1$ denote a class of distributions with finite mean $u_1(\mathrm{F}) = \mathbb{E}_\mathrm{F}[Y] < \infty$, for all $\mathrm{F} \in \mathcal{P}_1$, and let u_1 be the functional of interest, that is, $\mathrm{F} \equiv \mathrm{F}_\vartheta$, with $\vartheta = u_1(\mathrm{F}_\vartheta)$. The functional u_1 is elicitable with strictly consistent scoring function $\ell_1(\vartheta, y) = (\vartheta - y)^2$. The induced scoring rule $S_{u_1}(\mathrm{F}, y) = (\mathbb{E}_\mathrm{F}[Y] - y)^2$ is proper with respect to $\mathcal{P}_1$, but not strictly proper since distributions sharing the same mean yield the same expected score. This lack of strict propriety reflects the intentional localization of S_{u_1} to the mean, disregarding all other aspects of the distribution, and is therefore accommodated by the third condition of strict local propriety with respect to $(\mathcal{P}_1, u_1)$ in Definition 1. In this sense F_ϑ represents any distribution with mean ϑ, corresponding to the semi-parametric score-driven approach by Patton et al. (2019), albeit for a different functional.*

The expected score difference in Definition 1 satisfies the fundamental properties of a *local divergence* with respect to $\mathcal{T}$.

Definition 2. *Let $\mathcal{T} : \mathcal{P} \to \mathcal{V}$ be a functional. A map $\mathbb{D} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ is called a* local divergence with respect to $\mathcal{T}$ *if (i) $\mathcal{T}(\mathrm{F}) = \mathcal{T}(\mathrm{G})$ implies $\mathbb{D}(\mathrm{P}\|\mathrm{F}) = \mathbb{D}(\mathrm{P}\|\mathrm{G})$, $\forall \mathrm{P}, \mathrm{F}, \mathrm{G} \in \mathcal{P}$, (ii) $\mathbb{D}(\mathrm{P}\|\mathrm{F}) \geq 0$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$, and (iii) $\mathbb{D}(\mathrm{P}\|\mathrm{F}) = 0$ if and only if $\mathcal{T}(\mathrm{P}) = \mathcal{T}(\mathrm{F})$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$.*

The first condition ensures that the local divergence is insensitive to events outside the scope of $\mathcal{T}$. The second anchors the divergence in nonnegativity; no distribution can have

smaller divergence from P than P itself. The third ensures that only distributions equivalent to P under $\mathcal{T}$ have zero divergence from P, meaning that zero divergence uniquely identifies P up to $\mathcal{T}$-equivalence. For the censoring functional $v^{\flat}_{R_w}$, Definition 2 coincides with the notion of a local divergence introduced in Definition 2 of De Punder et al. (2025).

When considering expected score differences of any (potententially improper) scoring rule S, we follow Gneiting and Raftery (2007) by referring to them as *score divergences* $\mathbb{D}^{\dagger}_{S}(\mathrm{P}\|\mathrm{F}) := \mathbb{E}_{\mathrm{P}}S(\mathrm{F}, Y) - \mathbb{E}_{\mathrm{P}}S(\mathrm{P}, Y)$, where the superscript "†" emphasizes that the score divergence does not necessarily satisfy the conditions of a local divergence in Definition 2. For a meaningful comparison of distributions $\mathrm{F} \in \mathcal{P}$, the restriction $\mathbb{D}^{\dagger}_{S} = \mathbb{D}_{S}$ is *essential*, as without it, F could become statistically closer to P than P itself. With the restrictions in Definition 2 imposed, we refer to the triplet $(\mathcal{P}, \mathbb{D}, \mathcal{T})$ as a *local divergence space*.

3 PRADA

Pinning down the (infinite-dimensional) degree of freedom ϕ in an observation-driven time series model is a nontrivial task. Since Blasques et al. (2015), the class of score-driven models introduced by Creal et al. (2013) and Harvey (2013) has often been regarded as a solution to this problem, claiming its "information-theoretic optimality". As explained by De Punder et al. (2026), such a claim of *optimality*, in the sense of identifying an updating rule ϕ^* that yields maximal reduction in a KL-based divergence from the true distribution among all feasible updating rules ϕ, has never been shown; neither in Blasques et al. (2015) nor in De Punder et al. (2026). What was shown by De Punder et al. (2026), is that score-driven updates

$$\phi_{\mathrm{SD}}(y_t, \vartheta_{t|t-1}) \;\; := \vartheta_{t|t-1} - A\mathcal{S}_{t-1}s(y_t, \vartheta_{t|t-1}) \tag{1}$$

with $A\mathcal{S}_{t-1} \succ O_k$, where A is a parameter matrix and $\mathcal{S}_{t-1}$ an $\mathcal{F}_{t-1}$-measurable scaling matrix, are to a certain extent unique in ensuring EKL reductions for sufficiently small updates. So, when fixing the (local) divergence space to $(\mathcal{P}, \mathsf{KL}, \mathrm{id})$, where id is the identity functional, Theorem 1 of De Punder et al. (2026) identifies ϕ_{SD} (and equivalents thereof) as the class of updating rules for which sufficiently small updates guarantee improvements towards P_t in $(\mathcal{P}, \mathsf{KL}, \mathrm{id})$. The purpose of this section is to extend this characterization to more general local divergence spaces $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.

In the following, we first show that reductions can be guaranteed for any update implied by a scoring rule, as long as differences with respect to P_t are measured by the associated *score divergence*. However, as emphasized in Section 2.1 and in the discussion of the trimmed KL measure of Blasques et al. (2015) by De Punder et al. (2026), *only* score divergences based on strictly locally proper scoring rules induce local divergences and hence a meaningful comparison with the true distribution. PRADA updates are then defined as precisely those updates that guarantee reductions in local divergence space for sufficiently small updates.

3.1 Updates implied by scoring rules

We consider a generalization of score-driven updates ϕ_{SD} to the class of *derivative-adaptive* updating rules induced by a scoring rule S, that is,

$$\phi_S(y_t, \vartheta_{t|t-1}) := \vartheta_{t|t-1} - A_{t|t-1}\psi_S(y_t, \vartheta_{t|t-1}), \quad \psi_S(y_t, \vartheta_{t|t-1}) := \nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, y_t), \tag{2}$$

where $A_{t|t-1}$ denotes a $\mathcal{F}_{t-1}$-measurable learning rate matrix with $\|A_{t|t-1}\|_2 < \infty$. Unless stated otherwise, we assume $A_{t|t-1} \succ O_k$. The model-implied Hessian is similarly defined as $H_S(y_t, \vartheta_{t|t-1}) := \nabla^2_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, y_t)$. For $S = \mathrm{LogS}$, we obtain the score-driven update with $s(y_t, \vartheta_{t|t-1}) \equiv \psi_{\mathrm{LogS}}(y_t, \vartheta_{t|t-1})$, $H(y_t, \vartheta_{t|t-1}) \equiv H_{\mathrm{LogS}}(y_t, \vartheta_{t|t-1})$ and $A_{t|t-1} = A\mathcal{S}_{t|t-1}$.

Whether an updating rule belongs to the PRADA class will not depend on a more specific choice of $A_{t|t-1}$, paralleling that strict propriety is preserved under affine transformations of the scoring rule (Gneiting and Raftery 2007). The flexibility in specifying $A_{t|t-1}$ can be exploited, for instance, to minimize reductions in local divergence to the true distribution; see Dimitriadis et al. (2024) for a related discussion in the context of identification functions, where strict identification functions are shown to belong to similar equivalence classes.

Following Blasques et al. (2015) and De Punder et al. (2026), we evaluate the updating rule ϕ_S by the implied reduction of the associated score divergence from the true distribution P_t, under hypothetical redraws of $Y_t \sim \mathrm{P}_t$. Definition 3 formalizes this criterion, reducing to the EKL difference in Definition 2 of De Punder et al. (2026) for $S = \mathrm{LogS}$.

Definition 3. *Let Y_t and Y_t' denote independent copies with distribution P_t, $\phi : \mathcal{Y} \times \Theta \to \Theta$ an updating rule, and $\mathbb{D}_S^\dagger : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ a score divergence induced by the scoring rule S. The* expected score divergence difference *implied by $\mathbb{D}_S^\dagger$ and ϕ is defined as*

$$\Delta\mathbb{D}_S^\dagger(\phi) := \mathbb{E}_{Y_t \sim \mathrm{P}_t} \mathbb{D}_S^\dagger(\mathrm{P}_t \| \mathrm{F}_{t|t}) - \mathbb{D}_S^\dagger(\mathrm{P}_t \| \mathrm{F}_{t|t-1}),$$

where $\mathbb{D}_S^\dagger(\mathrm{P}_t \| \mathrm{F}_{t|t}) = \mathbb{E}_{Y_t' \sim \mathrm{P}_t}\left[S(\mathrm{F}_{t|t}, Y_t') - S(\mathrm{P}_t, Y_t')\right]$, in which the dependence on Y_t is implicit in $\mathrm{F}_{t|t} \equiv \mathrm{F}_{\vartheta_{t|t}}$, with $\vartheta_{t|t} = \phi(Y_t, \vartheta_{t|t-1})$.

A desirable property of an updating rule ϕ is to reduce the divergence of the model distribution from P_t under the update, that is, $\Delta\mathbb{D}_S^\dagger(\phi) < 0$, provided that $\mathbb{D}_S^\dagger$ is a local divergence (see Definition 2). The latter restriction is essential for drawing meaninful conclusions from reductions in $\mathbb{D}_S^\dagger$. Example 2 shows why. It considers an updating rule for the variance, evaluated by the score divergence induced by the *improper* Linear scoring rule, which is not a local divergence. In this case, reductions in the score divergence are entirely disconnected from the true distribution. Hence, decreases in this divergence are *uninformative* about the statistical distance with respect to the true distribution P_t.

Example 2. *Let $\mathcal{P}_\gamma$ be the class of zero mean Gaussian densities with variance γ. Given $y_t \in \mathbb{R}$ and $\gamma_{t|t-1} > 0$, $f_{t|t-1}$ is updated to $f_{t|t}$, with $\gamma_{t|t} = \phi_\gamma(y_t, \gamma_{t|t-1})$ for some updating rule ϕ_γ. The true variance is denoted γ_t. A direct calculation gives $\mathbb{E}_{p_t}\mathrm{LinS}(f_{t|t-1}, Y_t) = 1/\sqrt{2\pi(\gamma_t + \gamma_{t|t-1})}$. Correspondingly, $\Delta\mathbb{D}^\dagger_{\mathrm{LinS}}(\phi_\gamma) = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{\gamma_t+\gamma_{t|t-1}}} - \frac{1}{\sqrt{\gamma_t+\gamma_{t|t}}}\right) < 0$ if and only if $\gamma_{t|t} < \gamma_{t|t-1}$, for all γ_t. Therefore, a reduction toward the true density occurs as soon as the variance of $f_{t|t}$ is smaller than $f_{t|t-1}$, independent of what the actual variance is. Reductions in terms of $\mathbb{D}_{\mathrm{LinS}}$ are therefore completely uninformative about the relative divergence from the true density p_t. In contrast, $\Delta\mathbb{D}_{\mathrm{LogS}}(\phi_\gamma) < 0$ if and only if $\gamma_t > \log\left(\frac{\gamma_{t|t}}{\gamma_{t|t-1}}\right)\frac{\gamma_{t|t}\gamma_{t|t-1}}{\gamma_{t|t}-\gamma_{t|t-1}}$ for $\gamma_{t|t} > \gamma_{t|t-1}$, with the inequality reversed for $\gamma_{t|t} < \gamma_{t|t-1}$. Taking $\gamma_{t|t-1} = 1$ and $\gamma_{t|t} = 2$, a reduction in $\mathbb{D}_{\mathrm{LogS}}$ therefore only occurs if $\gamma_t > 2\log(2) \approx 1.38$, that is, if an update from $\gamma_{t|t-1}$ to $\gamma_{t|t}$ moves sufficiently toward the true variance γ_t.*

Below, we formalize that reductions in an expected score divergence can always be achieved through sufficiently small gradient updates implied by the scoring rule that defines the score divergence. In other words, the restriction to local divergences is not required for obtaining a negative $\Delta\mathbb{D}^\dagger_S(\phi)$, but only for ensuring that the reduction is informative. This observation supports the discussion of the trimmed KL (TKL) of Blasques et al. (2015) by De Punder et al. (2026), where it was argued that the results in Blasques et al. (2015) are formally correct yet uninformative because TKL is not a local divergence. To formally characterize the class of updating rules that reduce a score divergence, we first generalize the notion introduced by De Punder et al. (2026) of being *score equivalent in expectations* to that of being *expected gradient aligned*.

Definition 4. *An updating rule ϕ is said to be* P_t-expected gradient aligned (EGA) *for S at $\vartheta_{t|t-1} \in \Theta$ if $\left(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\right)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) < 0$, where $\Delta\phi(y, \vartheta) := \phi(y, \vartheta) - \vartheta$.*

The inner product of two expectations in Definition 4, rather than an expectation of

an inner product, intuitively follows from the definition of the expected score divergence $\mathbb{E}_{\mathrm{P}_t}\mathbb{D}_S^\dagger(\mathrm{P}_t\|\mathrm{F}_{t|t})$, in which the outer expectation averages over the hypothetical redraws of Y_t. By Theorem 1 below, reductions in an expected score divergence, i.e. $\Delta\mathbb{D}_S^\dagger(\phi) < 0$, are then guaranteed for sufficiently small updates, that is, for the *scaled* analogue ϕ_κ of ϕ.

Definition 5. *Let $\phi : \mathcal{Y} \times \Theta \to \Theta$ denote an updating rule, and $0 < \kappa < \infty$ a parameter. The scaled updating rule ϕ_κ of ϕ is such that $\vartheta_{t|t}^\kappa - \vartheta_{t|t-1} = \kappa(\vartheta_{t|t} - \vartheta_{t|t-1})$, where $\vartheta_{t|t}^\kappa := \phi_\kappa(y_t, \vartheta_{t|t-1})$.*

To establish an equivalence between updates that are EGA and reduce $\Delta\mathbb{D}_S^\dagger(\phi)$, we introduce, for any given $\vartheta_{t|t-1} \in \Theta$, the non-degeneracy condition for P_t:

$$\big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) \neq 0. \tag{ND}$$

The condition (ND) ensures that the first order effect in the expansion of the expected score divergence does not vanish. If $\Delta\phi(Y_t, \vartheta_{t|t-1}) = \alpha\psi_S(Y_t, \vartheta_{t|t-1})$, this condition rules out distributions P_t that render $\vartheta_{t|t-1}$ a pseudo-truth $\vartheta_{t|t-1}^*$, i.e. satisfying $\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}^*) = 0$. Given Assumption 1, 2 and 3 stated below, we then establish the result in Theorem 1.

Assumption 1. *The scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is twice continuously differentiable in ϑ for P_t-almost every $y \in \mathcal{Y}$, $\mathbb{E}_{\mathrm{P}_t}\|\Delta\phi(Y_t, \vartheta_{t|t-1})\|_2^2$ and $\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})$ are finite.*

Assumption 2. *The expected Hessian is uniformly bounded in the parameter ϑ, that is, $\sup_{\vartheta\in\Theta}\|\mathbb{E}_{\mathrm{P}_t}H_S(Y_t, \vartheta)\|_2 < \infty$.*

Assumption 3. *The expected Hessian is locally bounded in the parameter ϑ, that is, for any $\vartheta_{t|t-1} \in \Theta$, there exists a compact ball $\widetilde{B}_r \equiv \widetilde{B}_r(\vartheta_{t|t-1}) \subset \Theta$ of radius $r > 0$ around $\vartheta_{t|t-1}$ such that $\sup_{\vartheta\in\widetilde{B}_r(\vartheta_{t|t-1})}\|\mathbb{E}_{\mathrm{P}_t}H_S(Y_t, \vartheta)\|_2 < \infty$.*

Theorem 1. *Fix $\vartheta_{t|t-1} \in \Theta$, let $\mathrm{P}_t \in \mathcal{P}$ satisfy condition* (ND)*, and suppose Assumption 1 holds. If, additionally, Assumption 2 is satisfied, the following equivalence holds:*

$$\exists \bar{\kappa} > 0 \text{ such that } \forall \kappa \in (0, \bar{\kappa}] : \Delta \mathbb{D}_S^{\dagger}(\phi_\kappa) < 0 \iff \phi \text{ is EGA for } S \text{ at } \vartheta_{t|t-1}.$$

Moreover, this equivalence continues to hold under the relaxation of Assumption 2 to Assumption 3 for any bounded updated $\phi : \|\Delta\phi(Y_t, \vartheta_{t|t-1})\|_2 < d(\vartheta_{t|t-1}) < \infty$.

Theorem 1 characterizes the class of updating rules for which the integral $\Delta \mathbb{D}_S^{\dagger}(\phi_\kappa)$ becomes negative for sufficiently small $\kappa > 0$. If, *in addition*, S is strictly locally proper with respect to $(\mathcal{P}, \mathcal{T})$, this yields reduction within the local divergence space $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.

3.2 Proper updates

Building on the result of Theorem 1, we now refine this characterization by introducing a subclass of updating rules that guarantee movement toward the true distribution within some given local divergence space $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.

Definition 6. *Let $\mathcal{T} : \mathcal{P} \to \mathcal{V}$ denote a functional and let $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ be a scoring rule that is strictly locally proper with respect to $(\mathcal{P}, \mathcal{T})$. The updating rule $\phi : \mathcal{Y} \times \Theta \to \Theta$ is called* Proper and Robust Autoregressive Derivative Adaptive (PRADA) *with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$ if for every $\vartheta_{t|t-1} \in \Theta$, and every $\mathrm{P}_t \in \mathcal{P}$ satisfying* (ND)*, there exists $\bar{\kappa} > 0$ such that $\forall \kappa \in (0, \bar{\kappa}]$: $\Delta \mathbb{D}_S(\phi_\kappa) < 0$. The implied updates $\phi(y_t, \vartheta_{t|t-1})$ are called* PRADA updates *with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.*

Working on $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$, a canonical PRADA updating rule is the derivative-adaptive updating rule ϕ_S based on a scoring rule S that is strictly locally proper with respect to $(\mathcal{P}, \mathcal{T})$. This is formally stated in Corollary 1.

Corollary 1. *Consider a functional $\mathcal{T} : \mathcal{P} \to \mathcal{V}$. Let $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ be a scoring rule that is strictly locally proper with respect to $(\mathcal{P}, \mathcal{T})$. Assume $\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})$ to be non-zero and let Assumptions 1 and 2 hold. Then $\phi_S(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\psi_S(y_t, \vartheta_{t|t-1})$ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.*

Corollary 1 nests the result for score-driven updates in Corollary 1 of De Punder et al. (2026) as a special case. Specifically, Corollary 1 implies that ϕ_{LogS}, with $A_{t|t-1} = AD_{t|t-1} \succ O_k$, for some $\mathcal{F}_{t-1}$-measurable scaling matrix $D_{t|t-1}$, is a PRADA update with respect to $(\mathcal{P}, \mathsf{KL}, \mathrm{id})$. Moreover, by the characterizing part of Theorem 1, the PRADA update in Corollary 1 is, to a certain extent, unique. In particular, *only* updates for which $\mathbb{E}_{\mathrm{P}_t}[\psi_S(Y_t, \vartheta_{t|t-1})]^\top \big(\mathbb{E}_{\mathrm{P}_t}\phi(Y_t, \vartheta_{t|t-1}) - \vartheta_{t|t-1}\big) < 0$ can be PRADA updates with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$. Corollary 2 highlights an implication of this characterization.

Corollary 2. *Fix a triplet $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$, where $\mathbb{D}_S : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ denotes a local divergence with respect to $\mathcal{T}$. Consider the updating rule $\phi_\lambda(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\lambda(y_t, \vartheta_{t|t-1})$, for some measurable map $\lambda : \mathcal{Y} \times \Theta \to \mathbb{R}^k$. Assume $\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) \neq 0$, $\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1}) \neq 0$ and let Assumption 1 hold. Then ϕ_λ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$, if and only if*

$$\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1}) = a_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) + b_{t|t-1},$$

where $a_{t|t-1} > 0$ and $b_{t|t-1}$ is such that $\big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top A_{t|t-1}b_{t|t-1} = 0$, with $a_{t|t-1}$ and $b_{t|t-1}$ both $\mathcal{F}_{t-1}$-measurable.

As illustrated by Corollary 2, a specific local divergence space $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$ *necessitates* updates to align with the expected gradient of the strictly locally proper scoring rule S. This classification mirrors the equivalence between score divergences of strictly proper scoring rules and Bregman divergences advocated by Gneiting and Raftery (2007), with

canonical pairs such as LogS $\sim$ KL dating back to Good (1952). Hence, the characterization through expected $\mathbb{D}_S$ is considered natural, corresponding to an online version of *optimal score estimation* (Gneiting and Raftery 2007), for local scoring rules, or, more broadly, M-estimation. Generalizations of other performance criteria than EKL such as the conditional expected variation (CEV) of Gorgi et al. (2024) and expected GMM (EGMM) criterion of Creal et al. (2024) are therefore, albeit feasible, not considered; see De Punder et al. (2026) for a detailed comparison of these performance criteria.

The characterization of the class of PRADA updates in Corollary 2 confirms that the choice of the learning rate matrix is inconsequential for being a PRADA update or not. An analogous argument applies when including lagged terms of the gradients, i.e.,

$$\phi_S(y_t, \vartheta_{t|t-1}; \bar{p}) := \vartheta_{t|t-1} - \sum_{i=1}^{\bar{p}} \tilde{A}_{i,t|t-1} \psi_S(y_{t-i+1}, \vartheta_{t-i+1|t-i}). \tag{3}$$

Combining Equation (3) with the updating step $\vartheta_{t+1|t} = \omega + B\vartheta_{t|t}$, and adding lags gives, under reparameterization, the canonical prediction-to-prediction formulation of the PRADA($\bar{p}$,$\bar{q}$) model,

$$\vartheta_{t+1|t} := \omega - \sum_{i=1}^{\bar{p}} A_{i,t|t-1} \psi_S(y_{t-i+1}, \vartheta_{t-i+1|t-i}) + \sum_{j=1}^{\bar{q}} B_j \vartheta_{t-j+1|t-j},$$

which reduces to the well-known GAS($\bar{p}$,$\bar{q}$) specification of Creal et al. (2013, Eq. (2)) for the (negatively oriented) logarithmic scoring rule and $A_{i,t|t-1} = A_i D_{t|t-1}$, $i \in \{1, \dots \bar{p}\}$ for some coefficient vector ω and matrices $A_1, \dots, A_{\bar{p}}$, $B_1, \dots, B_{\bar{q}}$. For theory development, the arguments above allow us to focus on the canonical PRADA update in Corollary 1, by which we follow Gorgi (2020) and Blasques et al. (2023).

3.3 Robust updates

A well-known disadvantage of the logarithmic scoring rule, and, accordingly, score-driven updates, is their *unboundedness* (Selten 1998; Donker van Heel et al. 2025). We consider

an explicit example. Hereinafter, we occasionally omit the local divergence space when referring to PRADA updates when it is clear from the context.

Example 3. *Consider a Gaussian location-scale model with mean $\mu_{t|t-1}$ and variance $\sigma^2_{t|t-1} = \exp(\gamma_{t|t-1})$. Assume that $\mathbb{E}_{\mathrm{P}_t} Y_t^2 < \infty$. The model implies the score*

$$s(y_t, \vartheta_{t|t-1}) = -\frac{1}{\sigma^2_{t|t-1}} \begin{pmatrix} y_t - \mu_{t|t-1} \\ \frac{1}{2}\big((y_t - \mu_{t|t-1})^2 - \sigma^2_{t|t-1}\big) \end{pmatrix}$$

and Hessian

$$H(y_t, \vartheta_{t|t-1}) = \begin{pmatrix} \frac{1}{\sigma^2_{t|t-1}} & \frac{z_t}{\sigma_{t|t-1}} \\ \frac{z_t}{\sigma_{t|t-1}} & \frac{1}{2} z_t^2 \end{pmatrix},$$

where $z_t := (y_t - \mu_{t|t-1})/\sigma_{t|t-1}$. For a constant variance $\sigma^2_{t|t-1} = \exp(\gamma)$, $\forall t$, the expected Hessian is uniformly bounded in $\mu_{t|t-1}$ and the score-driven update is a PRADA update with respect to $(\mathbb{R}, \mathsf{KL}, \mathrm{id})$. In contrast, when fixing the mean $\mu_{t|t-1} = \mu$, $\forall t$, the expected Hessian is not uniformly bounded in γ_t, and therefore Assumption 2 of Corollary 1 not satisfied. Although the expected Hessian is locally bounded in the sense of Assumption 3, the update itself is unbounded. Therefore, the Assumption 3 part of Theorem 1 is not applicable, and EKL improvements cannot be guaranteed.

The sensitivity of the score-driven filter illustrated by Example 3 has been addressed in several ways. Blasques et al. (2023, p. 252) argue that the strong link between the model-implied conditional distribution and the updating equation in score-driven filters is not always desirable. However, by breaking this link, QSD updates, with $\tilde{s}(y_t, \vartheta_{t|t-1})$ for the quasi-score, constitute PRADA updates with respect to $(\mathbb{R}, \mathsf{KL}, \mathrm{id})$ if, and only if, $\mathbb{E}_{\mathrm{P}_t}\big[\widetilde{s}(Y_t, \vartheta_{t|t-1})\big]^\top A_{t|t-1} \mathbb{E}_{\mathrm{P}_t}\big[s(X_t, \vartheta_{t|t-1})\big] > 0$, which may, or may not hold, depending on the true distribution P_t (De Punder et al. 2026, App. F). Similarly, the QSD update in the

GARCH-T example of Blasques et al. (2023, p. 253) is a PRADA update, if and only if

$$\mathbb{E}_{\mathrm{P}_t}\left[Y_t^2 - \vartheta_{t|t-1}\right]\mathbb{E}_{\mathrm{P}_t}\left[\frac{\nu+1}{\nu-2+X_t^2/\vartheta_{t|t-1}}X_t^2 - \vartheta_{t|t-1}\right] > 0,$$

which holds trivially as $\nu \to \infty$, the score-driven case, but *not* in general.

We propose to retain the link between the model density and the updating rule while adapting the scoring rule instead. This can be accomplished by replacing the logarithmic scoring rule with a more robust alternative such as the CRPS (Catania et al. 2025), employing a weighted variant to *censor* or *winsorize* extreme observations (Harvey and Ito 2020), or by combining both approaches, for instance using the twCRPS of Gneiting and Ranjan (2011). Alternative scoring rules may also be desirable for reasons beyond robustness, analogous to their diverse use in forecast evaluation. In particular, the principle of coherent use of the same scoring rule for estimation and evaluation (Gneiting and Raftery 2007), extends naturally to its role in PRADA updates, inviting updates based on a rich variety of scoring rules. Detailed examples of unweighted and weighted scoring rules are deferred to Sections 3.4 and 3.5, respectively.

In the remainder of this section, we adopt the censoring framework of De Punder et al. (2025) to incorporate weight functions into PRADA updates. Censoring is appealing because it preserves the divergence $\mathbb{D}_S$, thereby directly identifying the local divergence space in which a censored updating rule is PRADA. This preservation stems from its projection-type nature, which we exploit to develop theory for unbounded updates such as those in Example 3. Although the associated Theorem 2 relies on indicator weight functions, we introduce censored PRADA updates for general weight functions in anticipation of Section 3.5.

For a regular scoring rule S, weight function w and nuisance distribution $\bar{\mathrm{H}} \in \mathcal{H} \subseteq \mathcal{P}$,

the *censored scoring rule* is defined as

$$S^\flat(\mathrm{F}_{t|t-1}, y_t) := w(y_t)S(\mathrm{F}^\flat_{t|t-1}, y_t) + \big(1 - w(y_t)\big)\mathbb{E}_{\bar{\mathrm{H}}}S(\mathrm{F}^\flat_{t|t-1}, Q_t), \tag{4}$$

where $\mathrm{F}^\flat_{t|t-1} \equiv v^\flat_w(\mathrm{F})$ is given by $\mathrm{dF}^\flat_{t|t-1} := \mathrm{dF}^w_{t|t-1} + \bar{F}_{t|t-1}\mathrm{d}\bar{\mathrm{H}}$, with $\mathrm{dF}^w_{t|t-1} := w\mathrm{dF}_{t|t-1}$ $\bar{F}_{t|t-1} := 1 - \int_{\mathcal{Y}} w\mathrm{dF}_{t|t-1}$ and $Q_t \sim \bar{\mathrm{H}}$, independently of Y_t. The weight function and nuisance distribution may be adapted to $\mathcal{F}_{t|t-1}$, although this is suppressed in the notation. The combination $S = \mathrm{LogS}$ and $\bar{\mathrm{H}} = \delta_r$, where δ_r denotes the Dirac measure at $r \in \mathcal{Y}$, implies the CSL of Diks et al. (2011). If the unweighted scoring rule is proper with respect to $\mathcal{P}^\flat := \{\mathrm{F}^\flat_{t|t-1}, \mathrm{F}_{t|t-1} \in \mathcal{P}\}$ then $S^\flat$ is strictly locally proper with respect to $(\mathcal{P}, v^\flat_{R_w})$, under the regularity given by Assumption 1 in De Punder et al. (2025). When no distribution $\mathrm{F}_{t|t-1} \in \mathcal{P}$ assigns positive mass to r, any weight function is compatible with choosing $\bar{\mathrm{H}} = \delta_r$ by this assumption.

The weight function and nuisance distribution are independent of $\vartheta_{t|t-1}$. Consequently,

$$\psi_{S^\flat}(y_t, \vartheta_{t|t-1}) = w(y_t)\nabla_\vartheta S(\mathrm{F}^\flat_{t|t-1}, y_t) + \big(1 - w(y_t)\big)\mathbb{E}_{\bar{\mathrm{H}}}\nabla_\vartheta S(\mathrm{F}^\flat_{t|t-1}, Q_t),$$

and a similar convex combination follows for the Hessian $H_{S^\flat}(y_t, \vartheta_{t|t-1})$. By Corollary 1,

$$\phi_{S^\flat}(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\psi_{S^\flat}(y_t, \vartheta_{t|t-1})$$

is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S^\flat}, v^\flat_{R_w})$, with $\mathbb{D}_{S^\flat}(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) = \mathbb{D}_S(\mathrm{P}^\flat_t \| \mathrm{F}^\flat_{t|t-1})$, under Assumptions 1, 2 and Assumption 1 in De Punder et al. (2025), and provided that $\bar{F}_{t|t-1} < 1$, for all $\mathrm{F}_{t|t-1} \in \mathcal{P}$.

To ensure bounded updates in the formal sense of Theorem 1, we consider censoring for the special case where $w(y_t) = \mathbb{1}_{\mathcal{C}}(y_t)$ and $\bar{\mathrm{H}} = \delta_r$. Then

$$S^\flat_{\mathcal{C}}(\mathrm{F}_{t|t-1}, y_t) := S\big(\mathrm{F}^\flat_{t|t-1}, Y^\flat_{\mathcal{C},t}(y_t)\big), \quad Y^\flat_{\mathcal{C},t}(y_t) := y_t\mathbb{1}_{\mathcal{C}}(y_t) + r\big(1 - \mathbb{1}_{\mathcal{C}}(y_t)\big), \tag{5}$$

is strictly locally proper for some $r \in \mathcal{Y}$ such that $\mathrm{F}_{t|t-1}(r) = 0$. Moreover, the censored distribution $\mathrm{F}^\flat_{\mathcal{C},t|t-1}$ is the pushforward measure of $\mathrm{F}_{t|t-1}$ by the $\mathcal{G}/\mathcal{G}^\flat_\mathcal{C}$-measurable map $Y^\flat_t : \mathcal{Y} \to \mathcal{Y}^\flat_\mathcal{C}$, where $\mathcal{Y}^\flat_\mathcal{C} := \mathcal{C} \cup \{r\}$ and $\mathcal{G}^\flat_\mathcal{C} := \sigma(\{E \cap \mathcal{C} : E \in \mathcal{G}\} \cup \{r\})$. The pushforward operator $\Pi_\mathcal{C}(\mathrm{F}) := Y^\flat_{\mathcal{C},t}\#\mathrm{F}$ is idempotent, as $Y^\flat_{\mathcal{C},t} \circ Y^\flat_{\mathcal{C},t} = Y^\flat_{\mathcal{C},t}$. Therefore $\Pi_\mathcal{C}$ acts as a projection from the space of distributions $\mathcal{P}$ on $(\mathcal{Y}, \mathcal{G})$ onto $\mathcal{P}^\flat_\mathcal{C} := \{\mathrm{F}^\flat_{\mathcal{C},t|t-1} : \mathrm{F}_{t|t-1} \in \mathcal{P}\}$ on $(\mathcal{Y}^\flat_\mathcal{C}, \mathcal{G}^\flat_\mathcal{C})$.

Updates that are already bounded allow for $\mathcal{C} = \mathcal{Y}$, in which case $\Pi_\mathcal{C} = \mathrm{id}$ and no censoring occurs. For unbounded updates, $\mathcal{C} \subset \mathcal{Y}$, and $\Pi_\mathcal{C}$ serves as a preliminary projection, ensuring boundedness. If $y \mapsto \psi_S(y, \vartheta_{t|t-1})$ is continuous for all $\vartheta_{t|t-1}$, then updates are trivially bounded on any compact $\mathcal{C}$ by Weierstrass. Since any compact subset suffices, one may take $\mathcal{C}$ large enough for $\Pi_\mathcal{C}$ to approximate the identity mapping in practice. In Example 3, for instance, while $(y_t - \mu_{t|t-1})^2$ is not uniformly bounded in y_t on $\mathbb{R}$, for any $\mu_{t|t-1} \in \mathbb{R}$, choosing $\mathcal{C} = [-\bar{c}, \bar{c}]$ with $\bar{c}$ equal to machine precision effectively renders $\Pi_\mathcal{C}$ redundant in applications. A formal result is stated in Theorem 2.

Assumption 4. *With $\mathcal{C} \subseteq \mathcal{Y}$ and $r \in \mathcal{Y}$ such that $\mathrm{F}(r) = 0, \forall \mathrm{F} \in \mathcal{P}$, the scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ and updating rule $\phi_S : \mathcal{Y} \times \Theta \to \Theta$ are such that for all $\vartheta_{t|t-1} \in \Theta$ there exists $d \equiv d(\vartheta_{t|t-1}) > 0 : \sup_{y \in \mathcal{C}} \|\psi_S(y, \vartheta_{t|t-1})\|_2 \vee \|\psi_S(r, \vartheta_{t|t-1})\|_2 < d$.*

Theorem 2. *Let $v : \mathcal{P} \to \mathcal{Q}$ be a functional, and $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ a scoring rule that is strictly locally proper with respect to $(\mathcal{P}, v)$. Let $\mathcal{C} \subseteq \mathcal{Y}$ be measurable, $r \in \mathcal{Y}$, and $S^\flat_\mathcal{C}$ defined in Equation (5), with $\mathrm{F}^\flat_{\mathcal{C},t|t-1} = Y^\flat_{\mathcal{C},t}\#\mathrm{F}_{t|t-1}$. Let Assumption 1, 3 and 4 hold, $\mathbb{E}_{\mathrm{P}_t}\psi_{S^\flat_\mathcal{C}}(Y_t, \vartheta_{t|t-1}) \neq 0$ and $\mathcal{P}^\flat_\mathcal{C} \subseteq \mathcal{P}$. Then $\phi_{S^\flat_\mathcal{C}}(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\psi_{S^\flat_\mathcal{C}}(y_t, \vartheta_{t|t-1})$ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S^\flat_\mathcal{C}}, v \circ \Pi_\mathcal{C})$, where $\mathbb{D}^\flat_{S_\mathcal{C}} = \mathbb{D}_S$ if $\Pi_\mathcal{C} = \mathrm{id}$.*

3.4 Examples for unweighted scoring rules

The class of PRADA updates accommodates a wide range of strictly proper unweighted scoring rules beyond the logarithmic score, which, as discussed in Section 3.3, is known for its sensitivity to outliers. Departing from the logarithmic scoring rule can therefore improve robustness, but also facilitates applicability to cases where only a distribution or characteristic function is available. For a comparison with score-driven updates, however, we generally assume the existence of a μ_t-density.

A family that nests the logarithmic scoring rule is the PseudoSpherical (PsSphS) family

$$\psi_{\mathrm{PsSphS}_\beta}(y_t, \vartheta_{t|t-1}) = -\frac{1}{\beta - 1}\left(\frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} - 1\right), \qquad \beta > 1,$$

where $\|f\|_\beta := (\int_{\mathcal{Y}} f^\beta \mathrm{d}\mu)^{1/\beta}$ denotes the L^β-norm on $(\mathcal{Y}, \mathcal{G})$, and the logarithmic scoring rule is obtained as $\beta \downarrow 1$. For all $\beta > 1$, the PseudoSpherical family is strictly proper with respect to the class of densities $\mathcal{P}_\beta$, satisfying $\|f\|_\beta < \infty$, for all $f \in \mathcal{P}_\beta$, and has divergence $\mathbb{D}_{\mathrm{PsSphS}_\beta}(p_t \| f_{t|t-1}) = \|p_t\|_\beta - \frac{\int_{\mathcal{Y}} p_t f_{t|t-1}^{\beta-1} \mathrm{d}\mu_t}{\|f_{t|t-1}\|_\beta^{\beta-1}}$ (Gneiting and Raftery 2007). The gradient reads

$$\psi_{\mathrm{PsSphS}_\beta}(y_t, \vartheta_{t|t-1}) = \frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}}\Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1})\Big),$$

which converges to $s(y_t, \vartheta_{t|t-1})$ as $\beta \downarrow 1$ since $\mathbb{E}_{f_{t|t-1}} s(Y_t, \vartheta_{t|t-1}) = 0$. The corresponding Hessian is provided in Appendix B.1, together with the analogous derivations for the Power scoring rule.

Example 4. *Consider the PseudoSpherical scoring rule with parameter $\beta > 1$ and reconsider the Gaussian location-scale model of Example 3. The implied gradient is*

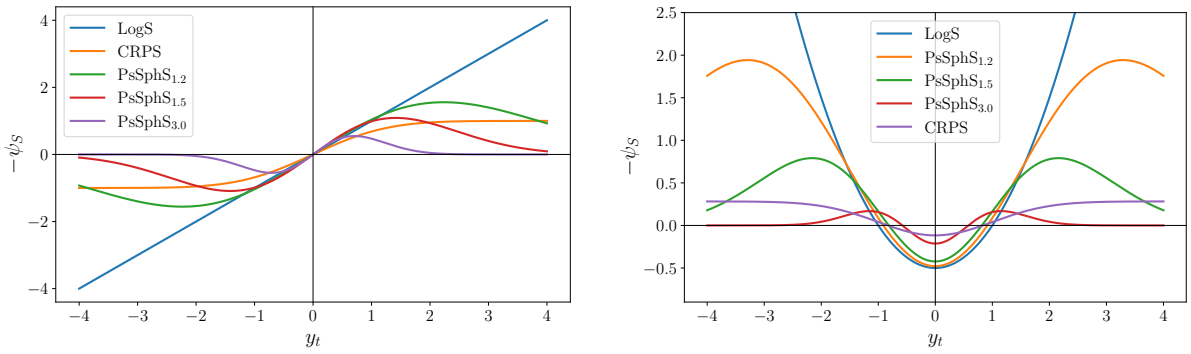
$$\psi_{\mathrm{PsSphS}_\beta}(y_t, \vartheta_{t|t-1}) = \left(\frac{\sigma_{t|t-1}}{\sqrt{2\pi/\beta}}\right)^{-\frac{\beta-1}{\beta}} \sqrt{2\pi}\varphi\left(\sqrt{\beta - 1}\, z_t\right) \begin{pmatrix} -\frac{z_t}{\sigma_{t|t-1}} \\ \frac{1}{2}(1 - z_t^2) - \frac{\beta-1}{2\beta} \end{pmatrix}.$$

For any fixed $\beta > 1$ and $\gamma_{t|t-1} \in \mathbb{R}$, the exponential factor ensures that both the gradient and Hessian are uniformly bounded in $y_t \in \mathbb{R}$. This contrasts with the logarithmic scoring rule

(obtained as $\beta \downarrow 1$), for which the gradient and Hessian are unbounded in the observation (see Example 3). The PseudoSpherical scoring rule with $\beta > 1$ therefore yields a robust alternative to the logarithmic scoring rule.

Figure 1 displays the marginal influence functions implied by the Gaussian location-scale model (fixing either location or scale) in Examples 3 and 4. The figure compares the PseudoSpherical scoring rule for several values of $\beta > 1$ with the logarithmic scoring rule and CRPS, the latter discussed later in Example 5. Whereas the logarithmic scoring rule produces unbounded responses for both mean and variance and is highly sensitive to outliers, the PseudoSpherical scoring rule yields smooth and bounded gradients, with robustness increasing in β.

Figure 1: Influence functions for the Gaussian location-scale model



(a) Mean $\mu_{t|t-1}$ (b) Log-variance $\gamma_{t|t-1}$

Influence function for the mean and log-variance in the Gaussian location-scale model based on the scoring rules LogS, PsSphS and CRPS, discussed in Examples 3, 4 and 5, respectively.

The Hyvärinen (2005) score provides another example closely related to the logarithmic scoring rule. For $\mathcal{Y} = \mathbb{R}^l$, it is defined as

$$\text{HS}(f_{t|t-1}, y_t) := \Delta_y \log f_{t|t-1}(y_t) + \frac{1}{2} \left\| \nabla_y \log f_{t|t-1}(y_t) \right\|_2^2,$$

where derivatives are now with respect to y_t. The Hyvärinen score is strictly proper with respect to the class of distributions whose densities $f_{t|t-1}(y)$ are twice continuously differentiable in y and satisfy $s(y, \vartheta_{t|t-1}) \to 0$, as $\|y\|_2 \to \infty$ (Parry et al. 2012, Sec. 2.5). Where the logarithmic scoring rule depends only on the density through the outcome, the Hyvärinen score only depends on the outcome through the first and second derivatives of the density; hence referred to as *local of order two* (Parry et al. 2012; Ehm and Gneiting 2012). The associated divergence reads $\mathbb{D}_{\text{HS}}(p_t \| f_{t|t-1}) = \frac{1}{2}\mathbb{E}_{p_t}\|\nabla_y \log p_t(Y_t) - \nabla_y \log f_{t|t-1}(Y_t)\|_2^2$, aiming to match the y-score $s_y(y_t, \vartheta_{t|t-1}) := \nabla_y \log f(y_t|\vartheta_{t|t-1})$ of $f_{t|t-1}$ with the y-score of the true density p_t.

The y-score matching objective of the Hyvärinen score is reflected by its gradient

$$\psi_{\text{HS}}(y_t, \vartheta_{t|t-1}) = \Delta_y s(y_t, \vartheta_{t|t-1}) + \big(\nabla_\vartheta s_y(y_t, \vartheta_{t|t-1})\big)^\top s_y(y_t, \vartheta_{t|t-1}), \tag{6}$$

which penalizes misalignment between the current y-score and the parameter gradient of the y-score. Importantly, the Laplacian term involves only second-order derivatives of the score with respect to $y \in \mathbb{R}^l$ but not the score itself, which also extends to the Hessian (see Appendix B.3). Because the Hyvärinen score is still local, now with respect to a neighborhood of y_t, rather than y_t itself, updates are expected to be more robust than those based on the logarithmic scoring rule, though not as robust as nonlocal scoring rules that depend on a global norm (PsSphS) or are distance sensitive (CRPS); see De Punder et al. (2025).

This general behavior is illustrated in the Gaussian location-scale model of Example 3, where the Hyvärinen score induces an inverse-variance scaling proportional to $1/\sigma^2_{t|t-1}$, dampening updates when the conditional variance is large. This enhances local robustness but does not eliminate unboundedness, aligning its properties more with the logarithmic score than with bounded alternatives such as the PsSphS and CRPS. Appendix E confirms these results by presenting the implied gradient and Hessian for the Gaussian univariate location-scale and the multivariate location model with fixed variance.

Leaving the class of density-based scoring rules, the rich class of *kernel scores* (Gneiting and Raftery 2007) generally remains applicable when only a distribution function is available. Their inherent distance-sensitivity, or at least, non-locality contributes to robustness in the resulting updates (De Punder et al. 2025). Let $\rho : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a negative definite kernel, defined as a symmetric function satisfying $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \rho(y_i, y_j) \leq 0$ for all $n \in \mathbb{N}$, $a_1, \ldots, a_n \in \mathbb{R}$. Then $S_\rho(\mathrm{F}_{t|t-1}, y_t) := \mathbb{E}_{\mathrm{F}_{t|t-1}}\rho(X_t, y_t) - \frac{1}{2}\mathbb{E}_{\mathrm{F}_{t|t-1}}\rho(X_t, X_t') - \frac{1}{2}\rho(y_t, y_t)$ defines a kernel score with associated score divergence $\mathbb{D}_{S_\rho}(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) = \mathbb{E}_{\mathrm{P}_t, \mathrm{F}_{t|t-1}}\rho(X_t, Y_t) - \frac{1}{2}\mathbb{E}_{\mathrm{P}_t}\rho(X_t, X_t') - \frac{1}{2}\mathbb{E}_{\mathrm{F}_{t|t-1}}\rho(Y_t, Y_t')$. The map $\mathbb{D}_{S_\rho}$ is a divergence if S_ρ is strictly proper, for which different assumptions exist (Gneiting and Raftery 2007; Steinwart and Ziegel 2021).

Under mild conditions allowing interchanging integration and differentiation, we obtain

$$\psi_{S_\rho}(y_t, \vartheta_{t|t-1}) = \mathbb{E}_{\mathrm{F}_{t|t-1}}\left[\left(\mathbb{E}_{\mathrm{F}_{t|t-1}}\rho(X_t, X_t') - \rho(X_t, y_t)\right) s(X_t, \vartheta_{t|t-1})\right], \tag{7}$$

with details (including H_{S_ρ}) deferred to Appendix B.4. Interestingly, the score operates within the expectation and is scaled by the realized ρ-similarity with regard to its expected value, both of which are helpful in robustifying the score-driven update. Indeed, while scores are unbounded for many model specifications, the expected score is often bounded. We consider an explicit example.

Example 5. *Consider the CRPS, that is, S_{ρ_1} with $\rho_1(y, y') = |y - y'|$. For the Gaussian case of Example 3, the CRPS simplifies to the expression provided by Gneiting and Raftery (2007, p.367), which implies the gradient and Hessian*

$$\nabla_\vartheta \mathrm{CRPS} = \begin{pmatrix} 1 - 2\Phi(z_t) \\ \sigma_{t|t-1}\tilde{\varphi}(z_t) \end{pmatrix}, \quad \nabla^2_\vartheta \mathrm{CRPS} = \begin{pmatrix} \frac{2z_t}{\sigma_{t|t-1}}\varphi(z_t) & z_t\varphi(z_t) \\ z_t\varphi(z_t) & 2\sigma_{t|t-1}\left(z_t^2\varphi(z_t) + \tilde{\varphi}(z_t)\right) \end{pmatrix},$$

where $z_t = (y_t - \mu_{t|t-1})/\sigma_{t|t-1}$, $\varphi(z) = 1/\sqrt{2\pi}\exp(-z^2/2)$, and $\tilde{\varphi}(z) = \varphi(z) - 1/(2\sqrt{\pi})$. For any fixed $\gamma_{t|t-1} \in \mathbb{R}$, both the gradient and Hessian are uniformly bounded in $y_t \in \mathbb{R}$ and $\mu_{t|t-1} \in \mathbb{R}$, since $\varphi(z)z^k$ is bounded $\forall k \in \mathbb{N}$. As shown in Figure 1, the CRPS displays winsorizing behavior in the mean component, smoothly approximating the Huber (1964) loss

in this case (Gneiting and Raftery 2007, Sec. 9.1). This is in contrast with LogS, for which the gradient and Hessian are unbounded in the observation.

3.5 Examples for weighted scoring rules

Section 3.3 introduced censoring as means to bound unweighted updates by applying a center indicator function. Robustification of updates can also be achieved through continuous weight functions, and the incorporation of other weight functions to emphasize different regions of interest follows their applications in focused forecast evaluation. A first example is the CSL by Diks et al. (2011), which is strictly locally proper with respect to $v^{\flat}_{R_w}$ (Holzmann and Klar 2017), so that

$$\psi_{\mathrm{LogS}^{\flat}_{w}}(y_t, \vartheta_{t|t-1}) = w(y_t)s(y_t, \vartheta_{t|t-1}) + \big(1 - w(y_t)\big)\mathbb{E}_{\mathrm{F}^{\sharp}_{1-w}} s(Y_t, \vartheta_{t|t-1}), \tag{8}$$

yields a PRADA update, with $\mathrm{dF}^{\sharp}_{w} := \frac{1-w}{\bar{F}_w}\mathrm{dF}_{t|t-1}$. The first term scales the score by the weight function, directly reducing the influence of extreme observations and allowing emphasis on specific regions of interest. The second term, which inclusion ensures strict local propriety, is generally bounded and therefore stabilizes the update when the raw score becomes unbounded.

Censoring can also be applied recursively: since Theorem 2 already admits strictly locally proper scoring rules such as the censored scoring rule itself, one may first emphasize a region such as the left tail of the distribution, and then apply a second round of censoring through the robustness operator $\Pi_{\mathcal{C}}$ as formalized in Theorem 2. For the Gaussian location-scale model introduced in Example 3, censoring the logarithmic score yields analytical expressions for both terms in Equation (8) from which the boundedness of $\mathbb{E}_{\mathrm{F}^{\sharp}_{1-w}} s(Y_t, \vartheta_{t|t-1})$ follows immediately (see Appendix E.2).

For general weight functions and unweighted scoring rules, De Punder et al. (2025,

App. C) show that Equation (4) admits the compact form

$$S^{\flat}_{w,\bar{\mathrm{H}}}(\mathrm{F}_{t|t-1}, y_t) = \mathbb{E}_{\mathrm{B}_{w(y_t)},\bar{\mathrm{H}}} S\big(\mathrm{F}^{\flat}_{t|t-1}, Y^{\flat}_t(y_t, Z_t, Q_t)\big),$$

accompanied with the censored random variable

$$Y^{\flat}_t(y_t, z_t, q_t) := \begin{cases} y_t, & \text{if } z_t = 1, \\ q_t, & \text{if } z_t = 0, \end{cases}$$

where $Z_t|(Y_t = y_t) \sim \mathrm{B}_{w(y_t)}$, the Bernoulli distribution with probability of success $w(y_t)$. Censored equivalents of gradients are then easily derived. For instance, the kernel score PRADA update with gradient (7) becomes

$$\psi_{S^{\flat}_{\rho}}(y_t, \vartheta_{t|t-1}) = \mathbb{E}_{\mathrm{B}_{w(y_t)},\bar{\mathrm{H}}} \mathbb{E}_{\mathrm{F}^{\flat}_{t|t-1}} \left[\left(\mathbb{E}_{\mathrm{F}^{\flat}_{t|t-1}} \big[\rho(X_t, X'_t)\big] - \rho(X_t, Y^{\flat}_t) \right) s(X_t, \vartheta_{t|t-1}) \right],$$

where the expectations over X_t, X'_t are with respect to $\mathrm{F}^{\flat}_{w,\bar{\mathrm{H}}}$, whereas the outer integral is with respect to $Z_t|(Y_t = y_t) \sim \mathrm{B}_{w(y_t)}$ and $Q_t \sim \bar{\mathrm{H}}$, conditional on the realization y_t. The same structure applies to the Hessian in Appendix B.4. For the nuisance distribution $\bar{\mathrm{H}}$, De Punder et al. (2025) recommends a convex combination of Dirac measures at the pivotal points of the weight function; e.g. $\bar{\mathrm{H}} = \gamma\delta_{r_1} + (1-\gamma)\delta_{r_2}$ for $w_1(y) = \mathbb{1}_{[r_1,r_2]}(y)$, with $\gamma \in [0,1]$.

As mentioned above, a key advantage of the censoring procedure is the preservation of the underlying divergence measure. Updating rules implied by $S^{\flat}_{w,\bar{\mathrm{H}}}$, which allow for PRADA updates with respect to $(\mathcal{P}, \mathbb{D}_{S^{\flat}_{w,\bar{\mathrm{H}}}} v^{\flat}_{R_w})$ by Corollary 1, are directly interpretable through the unweighted divergence $\mathbb{D}_S$. For example, the updates based on $\psi_{S^{\flat}_{\rho}}$, reduce the local divergence $\mathbb{D}_{S^{\flat}_{\rho}}(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) = \mathbb{E}_{\mathrm{P}^{\flat}_t, \mathrm{F}^{\flat}_{t|t-1}} \rho(X_t, Y_t) - \frac{1}{2}\mathbb{E}_{\mathrm{P}^{\flat}_t} \rho(X_t, X'_t) - \frac{1}{2}\mathbb{E}_{\mathrm{F}^{\flat}_{t|t-1}} \rho(Y_t, Y'_t)$, which follows directly from the expression for $\mathbb{D}_{S_\rho}$ given in Section 3.4.

Generalized censoring through Dirac measures is infeasible for the Hyvärinen score, which requires twice continuous differentiability. Instead, consider a continuous distribution $\bar{\mathrm{H}}$ on $\mathbb{R}^l$, with continuous density $\bar{h}$. Assumption 1 in De Punder et al. (2025) is satisfied if $\bar{\mathrm{H}}$

assigns positive probability to a measurable subset of the outcome space where $w(y)f(y) = 0$. Under this condition, the censored Hyvärinen score is strictly locally proper with respect to $v^{\flat}_{R_w}$. Hence, its gradient

$$\begin{aligned}
&\psi_{\mathrm{HS}^{\flat}_{w,\bar{h}}}(y_t, \vartheta_{t|t-1}) \\
&\quad = \bar{w}(y_t)\left(\Delta_y s(y_t, \vartheta_{t|t-1}) + \bar{w}(y_t)\nabla_y s(y_t, \vartheta_{t|t-1})\left(s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t)\right)\right) \\
&\qquad + \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big)\Xi(y_t, \vartheta_{t|t-1}),
\end{aligned}$$

facilitates a PRADA update, where

$$\Xi(y_t, \vartheta_{t|t-1}) \equiv \Xi\big(s_y(y_t, \vartheta_{t|t-1}), \nabla_y s(y_t, \vartheta_{t|t-1}), s^w_y(y_t), s^{\bar{h}}_y(y_t), d^{\sharp}_s(y_t, \vartheta_{t|t-1}), \bar{w}(y_t)\big),$$

$\bar{w}(y_t) := w(y_t)f_{t|t-1}(y_t)/f^{\flat}_{t|t-1}(y_t)$, $d^{\sharp}_s(y_t, \vartheta_{t|t-1}) := s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{f^{\sharp}_{t|t-1}} s(Y_t, \vartheta_{t|t-1})$, $s^w_y := \nabla_y \log w$ and $s^{\bar{h}}_y := \nabla_y \log \bar{h}$; see Appendix C.3 for further details.

The censored Hyvärinen scoring rule only depends on the score in deviation from the conditional expectation of the score. The latter is a consequence of the censoring procedure, similar to the additional dependence on $\bar{w}(y_t) \in (0, 1)$, $s^w_y(y_t)$ and $s^{\bar{h}}_y(y_t)$ relative to the unweighted gradient (6). The dependence on the score and $s^{\bar{h}}_y(y_t)$ vanish if $\bar{w}(y_t)\big(1 - \bar{w}(y_t)\big) = 0$ for all $y_t \in \mathbb{R}^l$. In that case, the censored Hyvärinen score depends on the nuisance density only through the weight function $\bar{w}(y_t)$. This condition would be trivially satisfied for indicator weight functions, but these are differentiable only for almost every $y_t \in \mathbb{R}^l$. Although the Hyvärinen score could, in principle, be extended to accommodate weak derivatives, it is often more convenient to use smooth approximations of indicators. An example is provided in Appendix E.3, illustrating how the contribution to the gradient through Ξ vanishes as the smooth approximation approaches the indicator function.

We next discuss some alternative localization methods. The threshold-weighted CRPS (twCRPS) of Gneiting and Ranjan (2011) is strictly locally proper with respect to $v^{\flat}_{w_1}$ for one-sided indicator functions (Holzmann and Klar 2017, Theorem 4) but not for $w_1(y_t) = \mathbb{1}_{R_1}(y_t)$, where $R_1 = [r_1, r_2]$, rendering the twCRPS non-localizing (Pelenis 2014).

Consequently $\mathbb{D}_{S^{\flat}_{\rho_1}}$ is not a local divergence with respect to $v^{\flat}_{w_1}$. As indicated by De Punder et al. (2025, App. E.4), the twCRPS can be formulated by the *winsorizing* functional

$$v^{\natural}_{w_1}\#\mathrm{F}_{t|t-1}(E) := \mathrm{F}_{t|t-1}(E \cap R_1) + \mathrm{F}_{t|t-1}\big((R_1^c)_1\big)\delta_{r_1}(E) + \mathrm{F}_{t|t-1}\big((R_1^c)_2\big)\delta_{r_2}(E),$$

$\forall E \in \mathcal{G}$, where $(R_1^c)_1 = (-\infty, r_1]$ and $(R_1^c)_2 = [r_2, \infty)$. Specifically,

$$\mathrm{twCRPS}_{w_1}(\mathrm{F}_{t|t-1}, y_t) = \mathbb{E}_{\mathrm{F}^{\natural}_{t|t-1}}\left|X_t - y_t^{\natural}\right| - \frac{1}{2}\mathbb{E}_{\mathrm{F}^{\natural}_{t|t-1}}\left|X_t - X_t'\right|,$$

where $y_t^{\natural} = y_t\mathbb{1}_{[r_1,r_2]}(y_t) + r_1\mathbb{1}_{(-\infty,r_1)}(y_t) + r_2\mathbb{1}_{(r_2,\infty)}(y_t)$ and $\mathrm{F}^{\natural}_{t|t-1} := v^{\natural}_{w_1}\#\mathrm{F}_{t|t-1}$, is strictly locally proper with respect to $(\mathcal{P}, v^{\natural}_{w_1})$. Rewriting the twCRPS in this way clarifies why it is non-localizing with respect to $v^{\flat}_{w_1}(\mathrm{F})$: equality $v^{\natural}_{w_1}(\mathrm{P}) = v^{\natural}_{w_1}(\mathrm{F})$ is not implied by $v^{\flat}_{w_1}(\mathrm{P}) = v^{\flat}_{w_1}(\mathrm{F})$ since the former requires knowledge of the individual $\mathrm{F}\big((R_1^c)_1\big)$ and $\mathrm{F}\big((R_1^c)_2\big)$.

A related example is the dynamic Tobit model of Harvey and Liao (2023), which is a score-driven model based on the winsorized likelihood

$$f^{\natural}_{w_1}(y_t) = f_{t|t-1}(y_t)^{\mathbb{1}_{[r_1,r_2]}(y_t)}F_{t|t-1}(r_1)^{\mathbb{1}_{(-\infty,r_1)}(y_t)}\big(1 - F_{t|t-1}(r_2)\big)^{\mathbb{1}_{(r_2,\infty)}(y_t)}, \quad y_t \in \mathbb{R},$$

corresponding to Equation (2) in Harvey and Liao (2023). The same reasoning extends to other scoring rules, that is, $S^{\natural}_{w_1} : \mathcal{P} \times \mathbb{R} \to \bar{\mathbb{R}}$, defined by $S(\mathrm{F}_{t|t-1}, y_t) := S(\mathrm{F}^{\natural}_{t|t-1}, y_t^{\natural})$ is strictly locally proper with respect to $v^{\natural}_{w_1}$, if $S : \mathcal{P}^{\natural}_{w_1} \times \mathbb{R} \to \bar{\mathbb{R}}$, where $\mathcal{P}^{\natural}_{w_1} := \{\mathrm{F}^{\natural}_{w_1}, \mathrm{F} \in \mathcal{P}\}$, denotes a scoring rule and $\mathrm{F}(R_1) < 1$, $\forall \mathrm{F} \in \mathcal{P}$, with $\mathcal{P}$ defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

While winsorizing arguably retains too much information of the distribution, conditioning may retain too little (De Punder et al. 2025). In particular, the conditional scoring rule $S_w(\mathrm{F}^{\sharp}_{t|t-1}, y_t) = w(y_t)S(\mathrm{F}^{\sharp}_{t|t-1}, y_t)$ is strictly locally proper with respect to the conditioning functional $v^{\sharp}_{R_w}\#\mathrm{F}(E) = \mathrm{F}(E)/\mathrm{F}(R_w)$, $\forall E \in \mathcal{G}^{\sharp}_{R_w} := \mathcal{G}|_{R_w}$, provided that $\mathrm{F}(R_w) > 0$. This result follows directly from Theorem 1 in Holzmann and Klar (2017).

In the context of kernel scores, Allen et al. (2023) also adopt a functional perspective by transforming unweighted kernel scores based on a kernel $\rho(y, y')$ through so-called chaining functions $\bar{u} : \mathcal{Y} \to \mathcal{Y}$. They characterize chaining functions for which the transformed

kernel score $\rho_{\bar{v}}(y, y') := \rho\big(\bar{u}(y), \bar{u}(y')\big)$ preserves strict local propriety relative to $\bar{v}$; see e.g., Propositions 4.3 to 4.5. However, constructing chaining functions that are both localizing and maintain strict propriety, proves difficult beyond the censoring transformation.

4 Implications for scoring functions

A particularly relevant subclass of the general results in Section 3 arises when the underlying scoring rule depends on the predictive distribution only through an elicitable functional, which may itself be adapted to $\mathcal{F}_{t-1}$, although this dependence is implicit in the notation. The class of PRADA updates encompasses such cases, corresponding to scoring functions $\ell : \mathcal{Y} \times \mathcal{Y} \to \bar{\mathbb{R}}$ that are strictly consistent with respect to some elicitable functional $u : \mathcal{P} \to \mathcal{Y}$. This follows from their equivalent formulation $S_\ell(\mathrm{F}_{t|t-1}, y_t) = \ell(u(\mathrm{F}_{t|t-1}), y_t)$, where strict consistency with respect to $(\mathcal{P}, u)$ is equivalent to strict local propriety with respect to $(\mathcal{P}, u)$.

A key implication of this formulation is that the scoring rule depends on $\mathrm{F}_{t|t-1}$ only through

$$\vartheta_{t|t-1} = u(\mathrm{F}_{t|t-1}). \tag{9}$$

Hence, no parametric specification is required for constructing PRADA updates derived from scoring functions (see Example 1). Appendix F clarifies why the strict consistency requirement in this argument is essential: replacing the scoring function by a loss that is not strictly consistent for the intended functional u may still yield reduction of the induced score divergence, but generally targets the corresponding Bayes-act functional rather than $u(\mathrm{P}_t)$. Moreover, given Equation (9), the link to strict consistency in Gneiting (2011) as discussed in Section 2.1 becomes immediate:

$$\mathbb{E}_{\mathrm{P}_t} S_\ell(\mathrm{F}_{t|t-1}, Y_t) - \mathbb{E}_{\mathrm{P}_t} S_\ell(\mathrm{P}_{t|t-1}, Y_t) = \mathbb{E}_{\mathrm{P}_t} \ell(\vartheta_{t|t-1}, Y_t) - \mathbb{E}_{\mathrm{P}_t} \ell(\vartheta_t, Y_t) \geq 0,$$

$\forall \vartheta_{t|t-1} \in \Theta$, with equality if and only if $\vartheta_{t|t-1} = u(\mathrm{P}_t) =: \vartheta_t$, $\forall \vartheta_{t|t-1} \in \Theta, \mathrm{P}_t \in \mathcal{P}$.

For the mean functional $u_1(\cdot) = \mathbb{E}_{(\cdot)}[Y_t]$ in Example 1, a PRADA update can be constructed without specifying $\mathrm{F}_{t|t-1}$, since all aspects of $\mathrm{F}_{t|t-1}$ other than its mean are absorbed by u_1. The corresponding scoring function $\ell_1(\vartheta, y) = (\vartheta - y)^2$, yields the gradient

$$\psi_{S_{\ell_1}}(y_t, \vartheta_{t|t-1}) = \nabla_\vartheta \ell_1(\vartheta_{t|t-1}, y_t) = -2(y_t - \vartheta_{t|t-1}),$$

which is proportional to the conditional mean update in the Gaussian location-scale model in Example 3 with fixed variance, where ℓ_1 appears in the exponent of the Gaussian likelihood.

This functional PRADA formulation nests classical models such as Autoregressive Moving Average (ARMA) for the conditional mean and the Autoregressive Conditional Heteroscedasticity (ARCH) and Generalized ARCH (GARCH) model for the conditional variance (Engle 1982; Bollerslev 1986). The next section discusses more recent examples of strictly consistent scoring functions.

4.1 Examples for scoring functions

In financial risk management, the VaR_β and Expected Shortfall (ES_β) functional are of particular interest due to the regulatory requirements set by the Third Basel Accord (Basel Committee, 2010). As demonstrated by Fissler and Ziegel (2016), the VaR_β and ES_β are jointly elicitable under the class of strictly consistent scoring functions defined in their Equation (5.5). Patton et al. (2019, Prop. 1) show that imposing the restriction

$$\ell_{\mathrm{FZ0},\beta}(\vartheta_{t|t-1}, y_t) := -\frac{\vartheta_{1,t|t-1} - y_t}{\beta \vartheta_{2,t|t-1}} \mathbb{1}_{\{y_t \leq \vartheta_{1,t|t-1}\}} + \frac{\vartheta_{1,t|t-1}}{\vartheta_{2,t|t-1}} + \log(-\vartheta_{2,t|t-1}) - 1$$

ensures that the loss difference is homogeneous of degree zero, provided that $\vartheta_{1,t|t-1} = \mathrm{VaR}_{\beta,t|t-1} < 0$ and $\vartheta_{2,t|t-1} = \mathrm{ES}_{\beta,t|t-1} < 0$, which is typically satisfied in risk management settings. Patton et al. (2019) then apply a score-driven modeling approach, for instance by assuming continuous distributions to avoid differentiability issues, to arrive at their updating equation for the VaR_β and ES_β tuple, which coincides with canonical PRADA(1,1)

update in Equation (3) implied by $\ell_{\mathrm{FZ0},\beta}$, with $A_{t|t-1} = A\tilde{H}_{t|t-1}^{-1}$, where $\tilde{H}_{t|t-1}$ denotes an $\mathcal{F}_{t-1}$-measurable scaling matrix related to the Hessian; see Equations (13) and (14) in Patton et al. (2019).

There also exist loss functions corresponding to interpolated functionals. An example from robust estimation is the Huber (1964) loss which is strictly consistent for its own functional, interpolating between the mean and the median. In analogy to $\ell_{\mathrm{FZ0},\beta}$, the Huber loss is twice continuously differentiable for P_t-almost every y_t for continuous distributions, thereby satisfying Assumption 1. Its derivative

$$\psi_{\mathrm{Huber},\kappa}(y_t, \vartheta_{t|t-1}) = -\begin{cases} y_t - \vartheta_{t|t-1}, & |y_t - \vartheta_{t|t-1}| \leq \kappa, \\ \kappa \operatorname{sgn}(y_t - \vartheta_{t|t-1}), & |y_t - \vartheta_{t|t-1}| > \kappa, \end{cases}$$

is Huber's (1964) classical ψ-function, which is clearly winsorizing. For a fixed variance, the CRPS-based update in Example 5 yields a smoothed version of $\psi_{\mathrm{Huber},\kappa}$, as illustrated in Figure 1a, aligning with the observation by Gneiting and Raftery (2007, Sec. 9.1). Smoothed analogues have also been proposed more generally. For instance, the Barron (2019) loss is infinitely differentiable in $\vartheta_{t|t-1}$ and its tuning parameters γ_1 and γ_2, with gradient

$$\psi_{\mathrm{Barron},\gamma_1,\gamma_2}(y_t, \vartheta_{t|t-1}) = \begin{cases} -\dfrac{(y_t - \vartheta_{t|t-1})}{\gamma_2^2}, & \gamma_1 = 2, \\ -\dfrac{2(y_t - \vartheta_{t|t-1})}{(y_t - \vartheta_{t|t-1})^2 + 2\gamma_2^2}, & \gamma_1 = 0, \\ -\dfrac{(y_t - \vartheta_{t|t-1})}{\gamma_2^2} \exp\Big(-\frac{1}{2}\left(\frac{y_t - \vartheta_{t|t-1}}{\gamma_2}\right)^2\Big), & \gamma_1 = -\infty, \\ -\dfrac{(y_t - \vartheta_{t|t-1})}{\gamma_2^2}\left(\frac{(y_t - \vartheta_{t|t-1})^2}{\gamma_2^2|\gamma_1 - 2|} + 1\right)^{\frac{\gamma_1}{2}-1}, & \text{otherwise.} \end{cases}$$

Similar to the Huber loss, the Barron loss nests consistent scoring functions for the mean ($\gamma_1 = 2$) and median ($\gamma_1 = 1$), and becomes equivalent to the local-mode finding loss of Van den Boomgaard and Van de Weijer (2002) for $\gamma_1 = -\infty$, though the mode itself is not elicitable (Heinrich 2014; Heinrich-Mertsching and Fissler 2022).

The Huber ψ arises naturally in the context of Z-estimation (Van der Vaart 1998, p. 43), the online analogue of which is discussed in the following section.

4.2 Updates based on identification functions

For parametric models, Fissler and Ziegel (2016) note that elicitability enables *M-estimation* (Huber 1964). Indeed, strict local propriety of S_ℓ with respect to $u(\mathrm{F}_{t|t-1}) = \vartheta_{t|t-1}$ mirrors to the classical identifiability assumption of the true parameter $\vartheta_t \in \Theta$ in M-estimation; see, e.g., Van der Vaart (1998, Theorem 5.7). Under sufficient smoothness conditions, the extremum problem of M-estimation corresponds to Z-estimation, e.g., $\mathbb{E}_{\mathrm{P}_t}[\vartheta_{t|t-1} - Y_t] = 0$ for the mean. The function $V_1(\vartheta, y) = (\vartheta - y)$ then serves as an *identification function* for the mean functional u_1 (Gneiting and Ranjan 2011; Fissler and Ziegel 2016), and $\mathbb{E}_{\mathrm{P}_t} V_1(\vartheta_{t|t-1}, Y_t) = 0$ constitutes a *moment condition*, as in the *generalized method of moments* framework of Hansen (1982).

Definition 7. *Consider a functional $u : \mathcal{P} \to \Theta$. A function $V : \Theta \times \mathcal{Y} \to \mathbb{R}^k$ is called a strict identification function with respect to $(\mathcal{P}, u)$ if $\mathbb{E}_{\mathrm{P}_t} V(\vartheta_{t|t-1}, Y_t) = 0$, if and only if $\vartheta_{t|t-1} = u(\mathrm{P}_t)$, for all $\mathrm{P}_t \in \mathcal{P}$.*

Identification functions are not unique. Any function $V_1 \in \{V_c(y, \vartheta) : c \in \mathbb{R}\backslash\{0\}\}$, where $V_{1,c}(y, \vartheta) := c(y - \vartheta)$ is also a strict identification function for the mean. According to Osband's principle for identification functions (Dimitriadis et al. 2024, Theorem 4), this set is complete. More generally, any strict identification $V(\vartheta, y)$ is unique only up to pre-multiplication with an invertible matrix-valued function $\bar{h} : \Theta \to \mathbb{R}^{k \times k}$, as established under the regularity conditions in Dimitriadis et al. (2024).

The restriction to *strict* identification functions relates to the identifiability assumption of the true parameter (functional) $\vartheta_t \in \Theta$ within Z-estimation (Van der Vaart 1998, Theorem 5.9). We now turn to an online analogue of Z-estimation, directly using the identification function. The circumvention of an explicit scoring function is feasible via Os-

band's principle (Fissler and Ziegel 2016, Theorem 3.2), for which we require the following regularity conditions.

Assumption 5. *The scoring function ℓ and identification function V satisfy*

(i) For every $\vartheta \in \mathrm{int}(\Theta)$, there are $P_{t,1}, \dots, P_{t,k+1} \in \mathcal{P}$ such that

$$0 \in \mathrm{int}\big(\mathrm{conv}\big(\mathbb{E}_{P_{t,1}} V(\vartheta, Y_t), \dots, \mathbb{E}_{P_{t,k+1}} V(\vartheta, Y_t)\big)\big).$$

(ii) For every $\mathrm{P}_t \in \mathcal{P}$, the function $\vartheta \mapsto \mathbb{E}_{\mathrm{P}_t} \ell(\vartheta, Y_t)$ is continuously differentiable on Θ.

(iii) Differentiation and integration can be interchanged, that is, $\nabla_\vartheta \mathbb{E}_{\mathrm{P}_t} \ell(\vartheta, Y_t) = \mathbb{E}_{\mathrm{P}_t} \nabla_\vartheta \ell(\vartheta, Y_t)$, for all $\vartheta \in \Theta$.

The first condition is a standard richness assumption corresponding to Assumption V1 in Fissler and Ziegel (2016) and Assumption 1 in Dimitriadis et al. (2024), and the second condition corresponds to Assumption S1 in Fissler and Ziegel (2016). The third condition is an additional smoothness condition to verify expected gradient alignment of the update.

Given Assumption 5, Osband's principle roughly states that a scoring function ℓ that is strictly consistent with respect to $(\mathcal{P}, u)$ and a function V that is a strict identification function with respect to $(\mathcal{P}, u)$ relate as

$$\mathbb{E}_{\mathrm{P}_t} \psi_{S_\ell}(Y_t, \vartheta_{t|t-1}) = h(\vartheta_{t|t-1}) \mathbb{E}_{\mathrm{P}_t} V(Y_t, \vartheta_{t|t-1}), \qquad \forall \vartheta_{t|t-1} \in \Theta, \tag{10}$$

for some matrix-valued function $h : \Theta \to \mathbb{R}^{k \times k}$. Under further smoothness conditions, Fissler and Ziegel (2016, Prop. 3.4) additionally establish a pointwise correspondence. However, the level of expected scores is already sufficient to obtain an equivalent condition for expected gradient alignment of an updating rule ϕ. Specifically, ϕ is EGA for S_ℓ at $\vartheta_{t|t-1} \in \Theta$ if and only if

$$\mathbb{E}_{\mathrm{P}_t}[V(Y_t, \vartheta_{t|t-1})]^\top h(\vartheta_{t|t-1})^\top \mathbb{E}_{\mathrm{P}_t} \Delta\phi(Y_t, \vartheta_{t|t-1}) < 0. \tag{11}$$

For elicitable and identifiable functionals u, with strict identification function V, the existence of a strictly consistent scoring function ℓ can therefore already be sufficient for a PRADA update, without knowledge of the explicit functional form of ℓ.

Corollary 3. *Consider a functional $u : \mathcal{P} \to \Theta$. Let $\ell : \Theta \times \mathcal{Y} \to \bar{\mathbb{R}}$ be a strictly consistent scoring function with respect to $(\mathcal{P}, u)$, and $V : \Theta \times \mathcal{Y} \to \mathbb{R}^k$ a strict identification function with respect to $(\mathcal{P}, u)$. Suppose that Assumption 1, 2 and 5 hold. Let $h : \Theta \to \mathbb{R}^{k \times k}$ be the function given in Equation (10). Assume $\mathbb{E}_{\mathrm{P}_t} V(\vartheta_{t|t-1}, Y_t) \neq 0$ and $h(\vartheta_{t|t-1}) \neq 0$. Then $\phi_V(y_t, \vartheta_{t|t-1}) := \vartheta_{t|t-1} - A_{t|t-1} V(\vartheta_{t|t-1}, y_t)$ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}, u)$ for all $A_{t|t-1}$ such that $h(\vartheta_{t|t-1})^\top A_{t|t-1} \succ O_k$, for all $\vartheta_{t|t-1} \in \Theta$.*

If $h(\vartheta_{t|t-1}) \succ (\prec)\ O_k$ for all $\vartheta_{t|t-1} \in \Theta$, the restriction on $A_{t|t-1}$ is satisfied, for instance, if $A_{t|t-1} = \alpha I_k$, where $\alpha > (<)\ 0$ and $A_{t|t-1} = A h(\vartheta_{t|t-1})$, where $A \succ O_k$. We give an example for each case, where $\mathcal{P}_1$ denotes a class of distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with strictly increasing and continuous distribution functions.

Example 6. *Consider the mean functional $u_1(\cdot) = \mathbb{E}_{(\cdot)} Y_t$ on $\mathcal{P}_1$, with strict identification function $V_1(\vartheta, y) = y - \vartheta$ and strict consistent scoring function $\ell_1(y, \vartheta) = (y - \vartheta)^2$. Correspondingly, $h(y) = 2$, and hence $\phi_{V_1}(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} + \alpha(y_t - \vartheta_{t|t-1})$ is a PRADA update with respect to $(\mathcal{P}_1, \mathbb{D}_{S_{\ell_1}}, u_1)$.*

Example 7. *For the quantile functional $u_{2,\beta}(\cdot) = q_\beta(\cdot)$, where $\beta \in (0, 1)$ on $\mathcal{P}_1$, the function $V_{2,\beta}(\vartheta, y) = \mathbb{1}_{\{y \leq \vartheta\}} - \beta$ is a strict identification function, and $\ell_{2,\beta}(y, \vartheta) = (y - \vartheta)\big(\beta - \mathbb{1}_{y \leq \vartheta}\big)$ a strictly consistent scoring function. Hence, Equation (10) holds with $h(\vartheta) = -1$ and $\phi_{V_{2,\beta}}(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} + \alpha\left(\mathbb{1}_{\{y_t \leq \vartheta\}} - \beta\right)$ is a PRADA update with respect to $(\mathcal{P}_1, \mathbb{D}_{S_{\ell_{2,\beta}}}, u_{2,\beta})$.*

Examples 6 and 7 illustrate that for univariate functionals, only the sign of $h(y)$ matters, which is implied by the chosen orientation of the identification function; e.g., the identification function $-V_{2,\beta}$ in Example 7 implies a positive $h(y)$. Hence, by fixing the

orientation of the identification function, we obtain an identification-based PRADA update where the specific h and ℓ are *implicit* by Osband's principle. For $d = 1$ and $k = 1$, an identification is oriented if $\mathbb{E}_{\mathrm{P}_t} V(\vartheta, Y_t) > 0$ if and only if $\vartheta > \vartheta_{t|t-1}$, for all $\vartheta_{t|t-1} \in \Theta$, $\vartheta \in \Theta$; see Fissler and Ziegel (2016, Rem. 4.1).

Proposition 1 formalizes such identification-based updates for a vector of univariate functionals, by which PRADA updates for the mean and β-quantiles follow directly. Proposition 1 considers the ratio of expectations functional. We first introduce some additional regularity, required for Propositions 4.2 and 4.4 of Fissler and Ziegel (2016), on which Propositions 1 and 2 are built, respectively.

Assumption 6. *For every $\mathrm{P}_t \in \mathcal{P}$, the function $\vartheta \mapsto \mathbb{E}_{\mathrm{P}_t} \ell(\vartheta, Y_t)$ is twice continuously differentiable on Θ with a gradient that is locally Lipschitz continuous, and $\vartheta \mapsto \mathbb{E}_{\mathrm{P}_t} V(\vartheta, Y_t)$ is continuously differentiable on Θ.*

Assumption 7. *For all $r \in \{1, \dots, k\}$ and for all $\vartheta_t \in \Theta$ there are $\mathrm{P}_t, \mathrm{P}_t' \in u^{-1}(\{\vartheta_t\})$ such that $\nabla_{\vartheta_j} \mathbb{E}_{\mathrm{P}_t} V_j(\vartheta_t, Y_t) = \nabla_{\vartheta_j} \mathbb{E}_{\mathrm{P}_t'} V_j(\vartheta_t, Y_t), \forall j \in \{1, \dots, k\} \backslash \{r\}$ and $\nabla_{\vartheta_r} \mathbb{E}_{\mathrm{P}_t} V_r(\vartheta_t, Y_t) \neq \nabla_{\vartheta_r} \mathbb{E}_{\mathrm{P}_t'} V_r(\vartheta_t, Y_t)$.*

Proposition 1. *Let $\Theta = \Theta_1 \times \ldots \times \Theta_k$, $u_m : \mathcal{P} \to \Theta_m \subseteq \mathbb{R}$ be surjective, elicitable and identifiable functionals, and $V_m : \Theta_m \times \mathcal{Y} \to \mathbb{R}$ the associated oriented strict identification function with respect to $(\mathcal{P}, u_m)$. Define $u = (u_1, \dots, u_m)$ and $V : \Theta \times \mathbb{R} \to \mathbb{R}^k$ by $V(\vartheta_{t|t-1}, y_t) = \big(V_1(\vartheta_{1,t|t-1}, y_t), \dots, V_k(\vartheta_{k,t|t-1}, y_t)\big)^\top$. Suppose that Assumptions 1, 2, 5, 6 and 7 hold. Then $\phi_V(y_t, \vartheta_{t|t-1}) := \vartheta_{t|t-1} - A_{t|t-1} V(\vartheta_{t|t-1}, y_t)$ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}, u)$ for all diagonal $A_{t|t-1} \succ O_k$, for some scoring function $\ell : \Theta \times \mathcal{Y} \to \bar{\mathbb{R}}$ that is strictly consistent with respect to $(\mathcal{P}, u)$.*

Proposition 2. *Let $u : \mathcal{P} \to \Theta$, be defined by $u(\cdot) := \mathbb{E}_{(\cdot)}[a(Y_t)]/\mathbb{E}_{(\cdot)}[b(Y_t)]$, where $a : \mathcal{Y} \to \mathbb{R}^k$ and $b : \mathcal{Y} \to \mathbb{R}$ are P_t-integrable functions, with $\mathbb{E}_{\mathrm{P}_t}[b(Y_t)] > 0$, for all $\mathrm{P}_t \in \mathcal{P}$. Define $V : \Theta \times \mathcal{Y} \to \mathbb{R}^k$, $V(\vartheta, y) = b(y)\vartheta - a(y)$. Moreover, let $\ell : \Theta \times \mathcal{Y} \to \mathbb{R}$ be a*

strictly consistent scoring function with respect to $(\mathcal{P}, u)$*. Suppose that Assumptions 1, 2, and 5 hold. Then* $\phi_V(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - \alpha V(\vartheta_{t|t-1}, y_t)$ *is a PRADA update with respect to* $(\mathcal{P}, \mathbb{D}_{S_\ell}, u)$ *for all* $\alpha > 0$*, for some scoring function* $\ell : \Theta \times \mathcal{Y} \to \bar{\mathbb{R}}$ *that is strictly consistent with respect to* $(\mathcal{P}, u)$*.*

For first order elicitable functionals such as the mean, Value-at-Risk and expectiles, Proposition 1 directly facilitates PRADA updates. For second order elicitable functionals such as those discussed by Dimitriadis et al. (2024), or more general multivariate functionals, the correspondence to identification functions is substantially more involved. As noted by Fissler and Ziegel (2016), such relations are best examined on a case-by-case basis.

5 Conclusion

This paper formalizes PRADA updates as the class of derivative-adaptive rules that, by construction, reduce a local divergence with respect to a prespecified functional. As a result, not all QSD updates are PRADA. QSD breaks the connection between the model-implied distribution and the updating direction, achieving divergence reductions only under additional alignment conditions that may not hold in practice. PRADA updates restore this link by tying the update to a strictly locally proper scoring rule or a strictly consistent scoring function, thereby ensuring movement toward the truth within a given local divergence space. PRADA updates accommodate robustness through weighted, for instance censored, scoring rules and include unweighted alternatives such as the PseudoSpherical, Hyvärinen, and kernel scores. These constructions retain interpretability because censoring preserves the associated unweighted divergence. Identification function representations further connect PRADA to online Z-estimation for elicitable functionals. With the introduction of PRADA updating rules, we provide a principled blueprint for dynamic updating: specify the target functional and its strictly proper or strictly consistent loss, derive the corresponding PRADA update, and inherit guaranteed divergence-reducing behavior.

Appendix

A Proofs of the main results

A.1 Proof of Theorem 1

Proof. We follow the structure of the argument in Appendix A of De Punder et al. (2026), which relies on the following exact Taylor expansion. For any fixed $y_1, y_2 \in \mathcal{Y}$, we apply the path-integral version of the Mean Value Theorem to $S(\mathrm{F}_{\vartheta^{\kappa}_{t|t}(y_2)}, y_1)$ at $\vartheta_{t|t-1}$ to obtain

$$\begin{aligned} S(\mathrm{F}_{\vartheta^{\kappa}_{t|t}(y_2)}, y_1) = S(\mathrm{F}_{t|t-1}, y_1) + \psi_S(y_1, \vartheta_{t|t-1})\big(\vartheta^{\kappa}_{t|t}(y_2) - \vartheta_{t|t-1}\big) \\ + \frac{1}{2}\big(\vartheta^{\kappa}_{t|t}(y_2) - \vartheta_{t|t-1}\big)^{\top} \tilde{H}_S(y_1, y_2, \vartheta_{t|t-1})\big(\vartheta^{\kappa}_{t|t}(y_2) - \vartheta_{t|t-1}\big), \end{aligned}$$

where $\widetilde{H}_S(y_1, y_2, \vartheta_{t|t-1}) := \int_0^1 (1-u) H_S\big(y_1, u\vartheta_{t|t-1} + (1-u)\vartheta^{\kappa}_{t|t}(y_2)\big)\mathrm{d}u$, with $\widetilde{\vartheta}^{\kappa}_u(y_2) := u\vartheta_{t|t-1} + (1-u)\vartheta^{\kappa}_{t|t}(y_2)$ is on the line between $\vartheta_{t|t-1}$ and $\vartheta^{\kappa}_{t|t}(y_2)$, for all $u \in [0,1]$ and $y_2 \in \mathcal{Y}$.

Then, by exploiting the independent copy assumption in Definition 3, we obtain

$$\begin{aligned} &\Delta\mathbb{D}^{\dagger}_S(\phi_\kappa) \\ &= \mathbb{E}_{\mathrm{P}_t}\mathbb{D}^{\dagger}_S(\mathrm{P}_t \| \mathrm{F}^{\kappa}_{t|t}) - \mathbb{E}_{\mathrm{P}_t}\mathbb{D}^{\dagger}_S(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) \\ &= \int_{\mathcal{Y}^2} \Big(S(\mathrm{F}_{\vartheta^{\kappa}_{t|t}(y_2)}, y_1) - S(\mathrm{F}_{t|t-1}, y_1) \Big) \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) \\ &= \int_{\mathcal{Y}^2} \psi_S(y_1, \vartheta_{t|t-1})^{\top} \big(\vartheta^{\kappa}_{t|t}(y_2) - \vartheta_{t|t-1}\big) \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) \\ &\quad + \frac{1}{2}\int_{\mathcal{Y}^2} \big(\vartheta^{\kappa}_{t|t}(y_2) - \vartheta_{t|t-1}\big)^{\top} \widetilde{H}_S(y_1, y_2, \vartheta_{t|t-1}) \big(\vartheta^{\kappa}_{t|t}(y_2) - \vartheta_{t|t-1}\big) \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) \\ &= \kappa \int_{\mathcal{Y}^2} \psi_S(y_1, \vartheta_{t|t-1})^{\top} \big(\phi(y_2, \vartheta_{t|t-1}) - \vartheta_{t|t-1}\big) \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) + \frac{\kappa^2}{2} I_1 \\ &= \kappa \left(\int_{\mathcal{Y}} \psi_S(y_1, \vartheta_{t|t-1}) \mathrm{P}_t(\mathrm{d}y_1) \right)^{\top} \int_{\mathcal{Y}} \Delta\phi(y_2, \vartheta_{t|t-1}) \mathrm{P}_t(\mathrm{d}y_2) + O(\kappa^2) \\ &= \kappa \mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t, \vartheta_{t|t-1})\big)^{\top} \mathbb{E}_{\mathrm{P}_t} \Delta\phi(Y_t, \vartheta_{t|t-1}) + O(\kappa^2), \end{aligned}$$

as $\kappa \downarrow 0$, where $\mathrm{P}_t^{\otimes 2} := \mathrm{P}_t \otimes \mathrm{P}_t$ denotes the product measure on $(\mathcal{Y}^2, \mathcal{G} \otimes \mathcal{G})$, where $\mathcal{Y}^2 := \mathcal{Y} \times \mathcal{Y}$, and

$$I_1 := \int_{\mathcal{Y}^2} \big(\Delta\phi(y_2, \vartheta_{t|t-1})\big)^\top \widetilde{H}_S(y_1, y_2, \vartheta_{t|t-1}) \Delta\phi(y_2, \vartheta_{t|t-1}) \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2).$$

The factorization of the product measure is verified as follows. First, Assumption 1 implies $\mathbb{E}_{\mathrm{P}_t} \big\|\Delta\phi(Y_t, \vartheta_{t|t-1})\big\|_2 < \infty$. Together with $\mathbb{E}_{\mathrm{P}_t} \big\|\Delta\phi(Y_t, \vartheta_{t|t-1})\big\|_2 < \infty$, which is also implied by Assumption 1, the integrand is absolutely integrable on $(\mathcal{Y}^2, \mathcal{G} \otimes \mathcal{G})$. Therefore, Fubini's theorem applies, and the product measure factorization follows.

We now invoke Assumption 2, by which there exists a constant $0 < c < \infty$ such that

$$\sup_{\vartheta \in \Theta} \int_{\mathcal{Y}} \|H_S(y_1, \vartheta)\|_2 \, \mathrm{P}_t(\mathrm{d}y_1) \leq c,$$

and hence

$$\begin{aligned}
&|I_1| \\
&= \left| \int_{\mathcal{Y}^2} \big(\phi(y_2, \vartheta_{t|t-1}) - \vartheta_{t|t-1}\big)^\top \widetilde{H}_S(y_1, y_2, \vartheta_{t|t-1}) \big(\phi(y_2, \vartheta_{t|t-1}) - \vartheta_{t|t-1}\big) \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) \right| \\
&\leq \int_{\mathcal{Y}^2} \left| \big(\Delta\phi(y_2, \vartheta_{t|t-1})\big)^\top \widetilde{H}_S(y_1, y_2, \vartheta_{t|t-1}) \Delta\phi(y_2, \vartheta_{t|t-1}) \right| \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) \\
&\leq \int_{\mathcal{Y}^2} \big\|\Delta\phi(y_2, \vartheta_{t|t-1})\big\|_2 \left\| \widetilde{H}_S(y_1, y_2, \vartheta_{t|t-1}) \Delta\phi(y_2, \vartheta_{t|t-1}) \right\|_2 \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) \\
&\leq \int_{\mathcal{Y}^2} \big\|\widetilde{H}_S(y_1, y_2, \vartheta_{t|t-1})\big\|_2 \, \big\|\Delta\phi(y_2, \vartheta_{t|t-1})\big\|_2^2 \, \mathrm{P}_t^{\otimes 2}(\mathrm{d}y_1, \mathrm{d}y_2) \\
&\leq \frac{1}{2} \sup_{\vartheta \in \Theta} \int_{\mathcal{Y}} \|H_S(y_1, \vartheta)\|_2 \, \mathrm{P}_t(\mathrm{d}y_1) \int_{\mathcal{Y}} \big\|\Delta\phi(y_2, \vartheta_{t|t-1})\big\|_2^2 \, \mathrm{P}_t(\mathrm{d}y_2) \\
&\leq \frac{c}{2} \mathbb{E}_{\mathrm{P}_t} \big\|\Delta\phi(Y_t, \vartheta_{t|t-1})\big\|_2^2 \\
&< \infty.
\end{aligned}$$

Here, the first inequality follows from the triangle inequality, the second inequality from Cauchy-Schwarz, the third inequality by definition of the operator norm, the fourth inequality from the triangle inequality, the linearity of the integral and a simplification origination

from the independent copy assumption in Definition 3, the fifth inequality by Assumption 2, and the last inequality by Assumption 1, specifically $\mathbb{E}_{\mathrm{P}_t}\|\Delta\phi(Y_t,\vartheta_{t|t-1})\|_2^2<\infty$.

If Assumption 2 does not hold, but Assumption 3 still holds, we verify the boundedness of I_1 for bounded updates. This can be verified in a spirit similar to the proof of Theorem 2 in Appendix A of De Punder et al. (2026). Specifically, fixing $\vartheta_{t|t-1}\in\Theta$, let $d\equiv d(\vartheta_{t|t-1})$ be a constant such that $\left\|\Delta\phi(Y_t,\vartheta_{t|t-1})\right\|_2<d$. Then, we can choose $\kappa\in(0,r/d)$ such that $\vartheta^{\kappa}_{t|t}(Y_t)\in\widetilde{B}_r(\vartheta_{t|t-1})$ a.s. and hence, by the convexity of $\widetilde{B}_r(\vartheta_{t|t-1})$ also $\widetilde{\vartheta}^{\kappa}_u=u\vartheta_{t|t-1}+(1-u)\vartheta^{\kappa}_{t|t}(y_2)\in\widetilde{B}_r(\vartheta_{t|t-1})$, for all $u\in[0,1]$. Therefore,

$$\begin{aligned}
|I_1| &\leq \frac{1}{2}\sup_{\vartheta\in\widetilde{B}_r(\vartheta_{t|t-1})}\int_{\mathcal{Y}}\|H_S(y_1,\vartheta)\|_2\,\mathrm{P}_t(\mathrm{d}y_1)\int_{\mathcal{Y}}\left\|\Delta\phi(y_2,\vartheta_{t|t-1})\right\|_2^2\,\mathrm{P}_t(\mathrm{d}y_2)\\
&\leq \frac{\tilde{c}}{2}\mathbb{E}_{\mathrm{P}_t}\left\|\Delta\phi(Y_t,\vartheta_{t|t-1})\right\|_2^2\\
&<\infty,
\end{aligned}$$

where $0<\tilde{c}<\infty$ is a constant such that $\sup_{\vartheta\in\widetilde{B}_r(\vartheta_{t|t-1})}\int_{\mathcal{Y}}\|H_S(y_1,\vartheta)\|_2\,\mathrm{P}_t(\mathrm{d}y_1)\leq\tilde{c}$, assumed to exist by Assumption 3. Hence, I_1 is also bounded for bounded updates under Assumption 3.

Following the proof of Theorem 1 in Appendix A of De Punder et al. (2026), we next consider the product of $\Delta\mathbb{D}_S^{\dagger}(\phi_\kappa)$ and $\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t,\vartheta_{t|t-1})\big)^{\top}\mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t,\vartheta_{t|t-1})$, that is,

$$\begin{aligned}
&\Delta\mathbb{D}_S^{\dagger}(\phi_\kappa)\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t,\vartheta_{t|t-1})\big)^{\top}\mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t,\vartheta_{t|t-1})\\
&=\kappa\Big(\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t,\vartheta_{t|t-1})\big)^{\top}\mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t,\vartheta_{t|t-1})\Big)^2+O(\kappa^2),
\end{aligned}\tag{A.1}$$

as $\kappa\downarrow 0$ since $\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t,\vartheta_{t|t-1})\big)^{\top}\mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t,\vartheta_{t|t-1})<\infty$ by assumption, given that $\mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t,\vartheta_{t|t-1})<\infty$ is implied by $\mathbb{E}_{\mathrm{P}_t}\|\Delta\phi(Y_t,\vartheta_{t|t-1})\|_2^2<\infty$.

To formally complete the proof, we start with the $\Longrightarrow$ direction. Suppose that there exists a $\bar{\kappa}>0$ such that $\Delta\mathbb{D}_S^{\dagger}(\phi_\kappa)<0$ for all $\kappa\in(0,\bar{\kappa}]$. Then, by Equation (A.1), there

exists a $\tilde{\kappa} > 0$ such that $0 < \tilde{\kappa} \leq \bar{\kappa}$:

$$\Delta\mathbb{D}_S^\dagger(\phi_{\tilde{\kappa}})\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) > 0$$

and hence $\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) < 0$ because $\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t, \vartheta_{t|t-1})\big)$ and $\mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1})$ are both non-zero by condition (ND), and $\Delta\mathbb{D}_S^\dagger(\phi_{\tilde{\kappa}}) < 0$, particularly for $\tilde{\kappa}$.

For the $\Longleftarrow$ direction, assume $\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) < 0$. By Equation (A.1), there exists a $\bar{\kappa} > 0$ such that

$$\Delta\mathbb{D}_S^\dagger(\phi_\kappa)\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) > 0,$$

for all $\kappa \in (0, \bar{\kappa}]$. Since $\mathbb{E}_{\mathrm{P}_t}\big(\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1})$ is assumed to be negative, it follows that $\Delta\mathbb{D}_S^\dagger(\phi_\kappa) < 0$ for all $\kappa \in (0, \bar{\kappa}]$. □

A.2 Proof of Corollary 1

Proof. The update $\phi_S(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\psi_S(y_t, \vartheta_{t|t-1})$ is EGA for S at $\vartheta_{t|t-1} \in \Theta$ since

$$\begin{aligned}&\big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi_S(Y_t, \vartheta_{t|t-1}) \\ &\quad = -\big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top A_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) < 0,\end{aligned}$$

using that $\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})$ is non-zero by assumption and $A_{t|t-1} \succ O_k$. Hence, by Theorem 1, specifically the " $\Longleftarrow$ " implication, there exists $\bar{\kappa} > 0$ such that for all $\kappa \in (0, \bar{\kappa}]$ it holds that $\Delta\mathbb{D}_S^\dagger(\phi_{S,\kappa}) < 0$, $\forall \vartheta_{t|t-1}, \mathrm{P}_t \in \mathcal{P}$. Moreover, since S is assumed to be strictly locally proper with respect to $(\mathcal{P}, \mathcal{T})$, the expected score difference $\mathbb{D}_S^\dagger(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) = \mathbb{E}_{\mathrm{P}_t}S(\mathrm{P}_t, Y_t) - \mathbb{E}_{\mathrm{P}_t}S(\mathrm{F}_{t|t-1}, Y_t)$ is a local divergence with respect to $(\mathcal{P}, \mathcal{T})$, i.e. $\mathbb{D}_S^\dagger = \mathbb{D}_S$. Therefore, the updating rule $\phi_S(y_t, \vartheta_{t|t-1})$ is PRADA with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.

A.3 Proof of Corollary 2

Consider the update $\phi_\lambda(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\lambda(y_t, \vartheta_{t|t-1})$. By Theorem 1, ϕ_λ is PRADA with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$ if, and only if,

$$\big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi_\lambda(Y_t, \vartheta_{t|t-1}) < 0,$$

that is, if and only if,

$$\big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top A_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1}) > 0.$$

Since $A_{t|t-1} \succ O_k$, $\langle x_1, x_2\rangle_{A_{t|t-1}} := x_1^\top A_{t|t-1}x_2$ defines an inner product on $\mathbb{R}^k$. Hence the inequality is equivalent to

$$\langle \mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}), \mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1})\rangle_{A_{t|t-1}} > 0.$$

By assumption, $\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) \neq 0$. Therefore, $\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1})$ must be a vector that lives in the subspace

$$\mathcal{H}^+_{A_{t|t-1}} := \Big\{x \in \mathbb{R}^k : \langle \mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}), x\rangle_{A_{t|t-1}} > 0\Big\}.$$

Since *every* vector $\tilde{x} \in \mathcal{H}^+_{A_{t|t-1}}$ admits the affine representation

$$\tilde{x} = a_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) + b_{t|t-1}$$

for some $\mathcal{F}_{t-1}$-measurable $a_{t|t-1} > 0$ and $b_{t|t-1} \in \mathbb{R}^k$ satisfying the $A_{t|t-1}$-orthogonality condition

$$\langle \mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}), b_{t|t-1}\rangle_{A_{t|t-1}} \equiv \big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top A_{t|t-1}b_{t|t-1} = 0,$$

the equivalence follows.

Formally, both directions of the if and only if can then be verified as follows. First, substituting $\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1}) = a_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) + b_{t|t-1}$ yields

$$\langle \mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}), \mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1})\rangle_{A_{t|t-1}} = a_{t|t-1}\|\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\|^2_{A_{t|t-1}} > 0,$$

where $\|x\|^2_{A_{t|t-1}} := \langle x, x\rangle_{A_{t|t-1}}$. Therefore every update with

$$\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1}) = a_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) + b_{t|t-1},$$

$a_{t|t-1} > 0$ and $\big(\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\big)^\top A_{t|t-1}b_{t|t-1} = 0$ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.

Conversely, suppose ϕ_λ is PRADA with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$. Then

$$\langle \mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}), \mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1})\rangle_{A_{t|t-1}} > 0,$$

and hence $\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1}) \in \mathcal{H}^+_{A_{t|t-1}}$. Write the orthogonal decomposition, that is, with respect to the inner product space $(\mathbb{R}^k, \langle\cdot,\cdot\rangle_{A_{t|t-1}})$, as

$$\mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1}) = a_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}) + b_{t|t-1},$$

where $\langle \mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}), b_{t|t-1}\rangle_{A_{t|t-1}} = 0$. Then

$$\langle \mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1}), \mathbb{E}_{\mathrm{P}_t}\lambda(Y_t, \vartheta_{t|t-1})\rangle_{A_{t|t-1}} = a_{t|t-1}\|\mathbb{E}_{\mathrm{P}_t}\psi_S(Y_t, \vartheta_{t|t-1})\|^2_{A_{t|t-1}} > 0,$$

which forces $a_{t|t-1} > 0$. Moreover, the $\mathcal{F}_{t-1}$-measurability of $a_{t|t-1}$ and $b_{t|t-1}$ follows from the measurability of $A_{t|t-1}$ and conditional expectations. □

A.4 Proof of Theorem 2

Proof. Let $\mathcal{C} \subseteq \mathcal{Y}$, $r \in \mathcal{Y}$ be such that Assumption 4 is satisfied. Fix a $\vartheta_{t|t-1} \in \Theta$ and let $d_1 = d_1(\vartheta_{t|t-1}) > 0$ be such that $\sup_{y\in\mathcal{C}} \left\|\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, y)\right\|_2 < d_1$. Moreover, let $d_2 \equiv d_2(\vartheta_{t|t-1}) > 0$ be such that $\left\|\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, r)\right\|_2 \leq d_2 < \infty$. Here, the constants $d_1, d_2 > 0$ exist by Assumption 4 since $\{\mathrm{F}^\flat_{\mathcal{C},t|t-1} : \mathrm{F}_{t|t-1} \in \mathcal{P}\} \subseteq \mathcal{P}$.

For all $y_t \in \mathcal{Y}$, the update

$$\begin{aligned}
&\phi_{S^\flat_{\mathcal{C}}}(y_t, \vartheta_{t|t-1}) \\
&= \vartheta_{t|t-1} - A_{t|t-1}\psi_{S^\flat_{\mathcal{C}}}(y_t, \vartheta_{t|t-1}) \\
&= \vartheta_{t|t-1} - A_{t|t-1}\left(\mathbb{1}_{\mathcal{C}}(y_t)\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, y_t) + \big(1 - \mathbb{1}_{\mathcal{C}}(y_t)\big)\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, r)\right),
\end{aligned}$$

yields

$$\begin{aligned}
&\left\|\Delta\phi_{S^\flat_{\mathcal{C}}}(y_t, \vartheta_{t|t-1})\right\|_2 \\
&= \left\|A_{t|t-1}\left(\mathbb{1}_{\mathcal{C}}(y_t)\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, y_t) + \big(1 - \mathbb{1}_{\mathcal{C}}(y_t)\big)\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, r)\right)\right\|_2 \\
&\leq \left\|A_{t|t-1}\right\|_2 \left\|\big(\mathbb{1}_{\mathcal{C}}(y_t)\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, y_t) + \big(1 - \mathbb{1}_{\mathcal{C}}(y_t)\big)\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, r)\big)\right\|_2 \\
&\leq \left\|A_{t|t-1}\right\|_2 \left(\mathbb{1}_{\mathcal{C}}(y_t)\left\|\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, y_t)\right\|_2 + \big(1 - \mathbb{1}_{\mathcal{C}}(y_t)\big)\left\|\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, r)\right\|_2\right) \\
&\leq \left\|A_{t|t-1}\right\|_2 \left(\sup_{y\in\mathcal{C}}\left\|\nabla_\vartheta S(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, y)\right\|_2 + d_2\right) \\
&\leq \left\|A_{t|t-1}\right\|_2 (d_1 + d_2) \\
&< \infty,
\end{aligned}$$

since $\left\|A_{t|t-1}\right\|_2 < \infty$ by assumption. Because $\left\|\Delta\phi_{S^\flat_{\mathcal{C}}}(y_t, \vartheta_{t|t-1})\right\|_2 < \infty$, it particularly follows that $\left\|\Delta\phi_{S^\flat_{\mathcal{C}}}(Y_t, \vartheta_{t|t-1})\right\|_2 < d_3$, for some $d_3 > 0$; i.e., the update $\phi_{S^\flat_{\mathcal{C}}}(Y_t, \vartheta_{t|t-1})$ is bounded.

Since $S : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is strictly locally proper with respect to $(\mathcal{P}, v)$,

$$\mathbb{D}_S(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) = \mathbb{E}_{\mathrm{P}_t} S(\mathrm{F}_{t|t-1}, Y_t) - \mathbb{E}_{\mathrm{P}_t} S(\mathrm{P}_t, Y_t) \geq 0,$$

with strict equality if and only if $v(\mathrm{P}_t) = v(\mathrm{F}_{t|t-1})$.

Applying the censoring procedure to this scoring rule then yields

$$S^\flat_{\mathcal{C}}(\mathrm{F}_{t|t-1}, y_t) = S\big(\mathrm{F}^\flat_{\mathcal{C},t|t-1}, Y^\flat_t(y_t)\big), \quad Y^\flat_t(y_t) := \begin{cases} y_t, & \text{if } y_t \in \mathcal{C}, \\ r, & \text{if } y_t \notin \mathcal{C}, \end{cases}$$

with $\mathrm{F}^\flat_{t|t-1,\mathcal{C}} = Y_t^\flat \# \mathrm{F}_{t|t-1}$ inducing the pushforward operator

$$\Pi_{\mathcal{C}} : \mathcal{P} \to \mathcal{P}, \quad \Pi_{\mathcal{C}}(\mathrm{F}) := Y_t^\flat \# \mathrm{F}.$$

Consequently,

$$\begin{aligned}
\mathbb{D}_{S^\flat_{\mathcal{C}}}(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) &= \mathbb{E}_{\mathrm{P}_t} S^\flat_{\mathcal{C}}(\mathrm{F}_{t|t-1}, Y_t) - \mathbb{E}_{\mathrm{P}_t} S^\flat_{\mathcal{C}}(\mathrm{P}_t, Y_t) \\
&= \mathbb{E}_{\mathrm{P}_t} S\big(\mathrm{F}^\flat_{t|t-1}, Y_t^\flat\big) - \mathbb{E}_{\mathrm{P}_t} S\big(\mathrm{P}^\flat_t, Y_t^\flat\big) \\
&= \mathbb{E}_{\mathrm{P}_t} S\big(\Pi_{\mathcal{C}}(\mathrm{F}_{t|t-1}), Y_t^\flat(Y_t)\big) - \mathbb{E}_{\mathrm{P}_t} S\big(\Pi_{\mathcal{C}}(\mathrm{P}_t), Y_t^\flat(Y_t)\big) \geq 0,
\end{aligned}$$

with strict equality if and only if $v(\mathrm{P}^\flat_t) = v(\mathrm{F}^\flat_{t|t-1})$, that is, if and only if

$$v \circ \Pi_{\mathcal{C}}(\mathrm{P}_t) = v \circ \Pi_{\mathcal{C}}(\mathrm{F}_{t|t-1}).$$

Hence, $S^\flat_{\mathcal{C}}$ is strictly locally proper with respect to $(\mathcal{P}, v \circ \Pi_{\mathcal{C}})$. Furthermore, $\mathbb{D}_{S^\flat_{\mathcal{C}}} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ is a local divergence with respect to $(\mathcal{P}, v \circ \Pi_{\mathcal{C}})$.

Moreover, since Assumption 1 and Assumption 3 hold, and $\phi_{S^\flat_{\mathcal{C}}}(y_t, \vartheta_{t|t-1})$ is bounded, it follows by Theorem 1 that there exists $\bar{\kappa} > 0$ such that for all $\kappa \in (0, \bar{\kappa}]$

$$\Delta \mathbb{D}_{S^\flat_{\mathcal{C}}}(\phi_{\kappa, S^\flat_{\mathcal{C}}}) < 0,$$

where $\phi_{\kappa, S^\flat_{\mathcal{C}}}$ is the linearly downscaled analogue of $\phi_{S^\flat_{\mathcal{C}}}$. Hence, $\phi_{S^\flat_{\mathcal{C}}}$ constitutes a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S^\flat_{\mathcal{C}}}, v \circ \Pi_{\mathcal{C}})$. For $\Pi_{\mathcal{C}} = \mathrm{id}$, we have $\mathcal{C} = \mathcal{Y}$, recovering $\mathrm{F}^\flat_{t|t-1,\mathcal{C}} = \mathrm{F}_{t|t-1}$ as the distribution of $Y^\flat_{\mathcal{C}_{t|t-1}} = Y_t$. Accordingly, $S^\flat_{\mathcal{C}} = S$ and $\mathbb{D}^\flat_{S_{\mathcal{C}}} = \mathbb{D}_S$. □

A.5 Proof of Corollary 3

First, we recall that Θ is assumed to be open and convex. Hence $\mathrm{int}(\Theta) = \Theta$ and Θ is a star domain. By Osband's principle (Fissler and Ziegel 2016, Thm. 3.2), there exists a matrix-valued function $h : \Theta \to \mathbb{R}^{k \times k}$ such that

$$\nabla_\vartheta \mathbb{E}_{\mathrm{P}_t} S_\ell(\vartheta_{t|t-1}, Y_t) = h(\vartheta_{t|t-1}) \mathbb{E}_{\mathrm{P}_t} V(\vartheta_{t|t-1}, Y_t), \tag{A.2}$$

for all $\vartheta_{t|t-1} \in \Theta$ and $\mathrm{P}_t \in \mathcal{P}$.

Under assumed interchangeability of integration and differentiation,

$$\mathbb{E}_{\mathrm{P}_t}\psi_{S_\ell}(Y_t, \vartheta_{t|t-1}) = h(\vartheta_{t|t-1})\mathbb{E}_{\mathrm{P}_t}V(\vartheta_{t|t-1}, Y_t).$$

Moreover, $\mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) = -A_{t|t-1}\mathbb{E}_{\mathrm{P}_t}V(\vartheta_{t|t-1}, Y_t)$ and hence

$$\begin{aligned}
&\big(\mathbb{E}_{\mathrm{P}_t}\psi_{S_\ell}(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi(Y_t, \vartheta_{t|t-1}) \\
&\quad = -\big(\mathbb{E}_{\mathrm{P}_t}V(\vartheta_{t|t-1}, Y_t)\big)^\top h(\vartheta_{t|t-1})^\top A_{t|t-1}\mathbb{E}_{\mathrm{P}_t}V(\vartheta_{t|t-1}, Y_t) \\
&\quad < 0,
\end{aligned}$$

since $h(\vartheta_{t|t-1})^\top A_{t|t-1} \succ O_k$ and $\mathbb{E}_{\mathrm{P}_t}V(\vartheta_{t|t-1}, Y_t) \neq 0$ by assumption.

Therefore, ϕ_V is EGA for S_ℓ at $\vartheta_{t|t-1}$, for all $\vartheta_{t|t-1} \in \Theta$. Combined with Assumption 1, expected gradient alignment is equivalent to strict reduction of the score divergence $\mathbb{D}^\dagger_{S_\ell}$ for a sufficiently small linear downscaling of the update, that is, by Theorem 1, there exists $\bar{\kappa} > 0$ such that for all $\kappa \in (0, \bar{\kappa}]$: $\Delta\mathbb{D}_{S_\ell}(\phi_{V,\kappa}) < 0$, where $\phi_{V,\kappa}$ denotes the linearly downscaled version of ψ_V. Additionally, since ℓ is strictly consistent with respect to $(\mathcal{P}, u)$, the scoring rule S_ℓ is strictly locally proper with respect to $(\mathcal{P}, u)$. Hence, $\mathbb{D}^\dagger_S = \mathbb{D}_S$, and ψ_V is PRADA with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}, u)$.

A.6 Proof of Proposition 1

First, recall that Θ is a star domain since Θ is assumed to be open and convex. Define $\Theta'_m := \{\vartheta_m : \exists(\vartheta'_1, \ldots, \vartheta'_k) \in \Theta, \vartheta'_m = \vartheta_m\}$. Then, by Proposition 4.2 in Fissler and Ziegel (2016) there exist functions $h'_m : \Theta'_m \to \mathbb{R}_{++} := (0, \infty)$ such that $h_{mm}(\vartheta_{t|t-1}) = h'_m(\vartheta_{m,t|t-1})$ for all $m \in \{1, \ldots, k\}$ and $\vartheta_{t|t-1} \in \Theta$, and $h_{rl}(\vartheta_{t|t-1}) = 0$ for all $r, l \in \{1, \ldots, k\}$, $l \neq r$, and for all $\vartheta_{t|t-1} \in \Theta$. In other words, Equation (A.2) is satisfied for a function $h : \Theta \to \mathbb{R}^{k\times k}$ that is positive definite and diagonal for all $\vartheta_{t|t-1} \in \Theta$.

Since $h(\vartheta_{t|t-1}) \succ O_k$ and diagonal for all $\vartheta_{t|t-1} \in \Theta$, it follows that $h(\vartheta_{t|t-1})^\top A_{t|t-1} \succ O_k$ for all diagonal $A_{t|t-1} \succ O_k$. Hence, ϕ_V is PRADA with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}, u)$ by Corollary 3, in which ℓ is a strictly consistent scoring function satisfying

$$\mathbb{E}_{\mathrm{P}_t} \psi_{S_\ell}(Y_t, \vartheta_{t|t-1}) = h(\vartheta_{t|t-1}) \mathbb{E}_{\mathrm{P}_t} V(\vartheta_{t|t-1}, Y_t).$$

from which it also more directly follows that ϕ_V is PRADA, since

$$\begin{aligned} &\big(\mathbb{E}_{\mathrm{P}_t} \psi_{S_\ell}(Y_t, \vartheta_{t|t-1})\big)^\top \mathbb{E}_{\mathrm{P}_t} \Delta\phi_V(Y_t, \vartheta_{t|t-1}) \\ &\quad = -\sum_{i=1}^{k} h_i'(\vartheta_{i,t|t-1}) [A_{t|t-1}]_{ii} \left[\mathbb{E}_{\mathrm{P}_t} V(\vartheta_{t|t-1}, Y_t)\right]_i^2 < 0. \end{aligned}$$

A.7 Proof of Proposition 2

Recall that Θ is a star shaped. By Proposition 4.4 of Fissler and Ziegel (2016), it immediately follows that the matrix $h(\vartheta_{t|t-1})$ in Equation (A.2) positive definite for all $\vartheta_{t|t-1} \in \Theta$. Hence, $\alpha h(\vartheta_{t|t-1}) \succ O_k$ for all $\alpha > 0$. Consequently, Corollary 3 implies that ϕ_V is PRADA with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}, u)$, in which ℓ is a strictly consistent scoring function satisfying

$$\mathbb{E}_{\mathrm{P}_t} \psi_{S_\ell}(Y_t, \vartheta_{t|t-1}) = h(\vartheta_{t|t-1}) \mathbb{E}_{\mathrm{P}_t} V(\vartheta_{t|t-1}, Y_t).$$

B Derivations Section 3.4

B.1 Density-based updates

In the derivations below, we will use the following result

$$\begin{aligned}\nabla_\vartheta \log \|f_{t|t-1}\|_\beta &= \nabla_\vartheta \log \left(\left(\int_{\mathcal{Y}} f_{t|t-1}(y)^\beta \mu(\mathrm{d}y) \right)^{1/\beta} \right) \\ &= \frac{1}{\beta} \frac{1}{\|f_{t|t-1}\|_\beta^\beta} \int_{\mathcal{Y}} \nabla_\vartheta f_{t|t-1}(y)^\beta \mu(\mathrm{d}y) \\ &= -\frac{1}{\|f_{t|t-1}\|_\beta^\beta} \int_{\mathcal{Y}} f(y|\vartheta_{t|t-1})^{\beta-1} s(y, \vartheta_{t|t-1}) f(y|\vartheta_{t|t-1}) \mu(\mathrm{d}y) \\ &= -\mathbb{E}_{\tilde{f}^\beta_{t|t-1}} s(Y_t, \vartheta_{t|t-1}), \end{aligned} \tag{B.1}$$

where $\tilde{f}^\beta_{t|t-1} := f^\beta_{t|t-1} / \|f_{t|t-1}\|^\beta_\beta$ defines a μ_t-density on $(\mathcal{Y}, \mathcal{G})$. Then

$$\begin{aligned}&\nabla_\vartheta \mathrm{PsSphS}_\beta(f_{t|t-1}, y_t) \\ &= -\frac{1}{\beta-1} \nabla_\vartheta \exp\left((\beta-1) \log f(y_t|\vartheta_{t|t-1}) - (\beta-1) \log \|f_{t|t-1}\|_\beta \right) \\ &= -\frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} \left(\nabla_\vartheta \log f(y_t|\vartheta_{t|t-1}) - \nabla_\vartheta \log \|f_{t|t-1}\|_\beta \right) \\ &= \frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} \Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}^\beta_{t|t-1}} s(Y_t, \vartheta_{t|t-1}) \Big) \\ &= \frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} \Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}^\beta_{t|t-1}} s(Y_t, \vartheta_{t|t-1}) \Big). \end{aligned}$$

For the Hessian, we note that

$$\frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} = -(\beta-1)\mathrm{PsSphS}_\beta(f_{t|t-1}, y_t) + 1$$

and therefore

$$\frac{\partial}{\partial \vartheta_{t|t-1}^\top} \frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} \tag{B.2}$$

$$= -(\beta-1) \frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} \Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}^\beta_{t|t-1}} s(Y_t, \vartheta_{t|t-1}) \Big)^\top. \tag{B.3}$$

Moreover,

$$
\begin{aligned}
&\frac{\partial}{\partial \vartheta_{t|t-1}^\top} \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) \\
&\quad = \int_{\mathcal{Y}} \nabla_\vartheta s(y_t, \vartheta_{t|t-1}) \frac{f_{t|t-1}(y_t)^\beta}{\|f_{t|t-1}\|_\beta^\beta} \mu(\mathrm{d}y_t) \\
&\quad = \int_{\mathcal{Y}} H(y_t, \vartheta_{t|t-1}) \frac{f_{t|t-1}(y_t)^\beta}{\|f_{t|t-1}\|_\beta^\beta} + s(y_t, \vartheta_{t|t-1}) \frac{\partial}{\partial \vartheta_{t|t-1}^\top} \frac{f_{t|t-1}(y_t)^\beta}{\|f_{t|t-1}\|_\beta^\beta} \mu(\mathrm{d}y_t) \\
&\quad = \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t, \vartheta_{t|t-1}) \\
&\qquad - \beta \int_{\mathcal{Y}} s(y_t, \vartheta_{t|t-1}) \frac{f_{t|t-1}(y_t)^\beta}{\|f_{t|t-1}\|_\beta^\beta} \Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) \Big)^\top \mu(\mathrm{d}y_t) \\
&\quad = \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t, \vartheta_{t|t-1}) \\
&\qquad - \beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} \left[s(Y_t, \vartheta_{t|t-1}) \Big(s(Y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) \Big)^\top \right] \\
&\quad = \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t, \vartheta_{t|t-1}) \\
&\qquad - \beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} \left[s(Y_t, \vartheta_{t|t-1}) s(Y_t, \vartheta_{t|t-1})^\top \right] \\
&\qquad - \beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} \left[s(Y_t, \vartheta_{t|t-1}) \right] \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} \left[s(Y_t, \vartheta_{t|t-1}) \right]^\top \\
&\quad = \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t, \vartheta_{t|t-1}) - \beta \mathbb{C}\mathrm{ov}_{\tilde{f}_{t|t-1}^\beta} (s(Y_t, \vartheta_{t|t-1})).
\end{aligned}
$$

Applying the product rule, we obtain

$$
\begin{aligned}
&\nabla_\vartheta^2 \mathrm{PsSphS}_\beta(f_{t|t-1}, y_t) \\
&= \frac{f_{t|t-1}(y_t)^{\beta-1}}{\|f_{t|t-1}\|_\beta^{\beta-1}} \Big(H(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t, \vartheta_{t|t-1}) + \beta \mathbb{C}\mathrm{ov}_{\tilde{f}_{t|t-1}^\beta} (s(Y_t, \vartheta_{t|t-1})) \\
&\quad - (\beta - 1) \Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) \Big) \Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) \Big)^\top \Big).
\end{aligned}
$$

As anticipated,

$$
\lim_{\beta \downarrow 1} \nabla_\vartheta^2 \mathrm{PsSphS}_\beta(f_{t|t-1}, y_t) = H(y_t, \vartheta_{t|t-1})
$$

because $\mathbb{E}_{f_{t|t-1}} s(Y_t, \vartheta_{t|t-1}) = 0$ and the information matrix equality for the negatively oriented logarithmic scoring rule holds, that is, $\mathbb{E}_{f_{t|t-1}} H(Y_t, \vartheta_{t|t-1}) = \mathbb{C}\mathrm{ov}_{f_{t|t-1}} (s(Y_t, \vartheta_{t|t-1}))$.

For the Power scoring rule

$$\mathrm{PowS}_\beta(f_{t|t-1}, y_t) = -\frac{\beta}{\beta-1}\left(f_{t|t-1}(y_t)^{\beta-1} - (\beta-1)\|f_{t|t-1}\|_\beta^\beta - 1\right), \quad \beta > 1,$$

we recall Equation (B.1) and note

$$\nabla_\vartheta \|f_{t|t-1}\|_\beta^\beta = \nabla_\vartheta \exp\left(\beta \log \|f_{t|t-1}\|_\beta\right) = -\beta\|f_{t|t-1}\|_\beta^\beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}).$$

Correspondingly,

$$\nabla_\vartheta \mathrm{PowS}_\beta(f_{t|t-1}, y_t) = \beta f_{t|t-1}(y_t)^{\beta-1} s(y_t, \vartheta_{t|t-1}) - \beta^2 \|f_{t|t-1}\|_\beta^\beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}),$$

for which $\lim_{\beta\downarrow 1} \nabla_\vartheta \mathrm{PowS}_\beta(f_{t|t-1}, y_t) = s(y_t, \vartheta_{t|t-1})$.

For the Hessian, we observe that

$$\begin{aligned}
&\frac{\partial}{\partial \vartheta_{t|t-1}^\top} f_{t|t-1}(y_t)^{\beta-1} s(y_t, \vartheta_{t|t-1}) \\
&\quad = f_{t|t-1}(y_t)^{\beta-1}\left(H(y_t, \vartheta_{t|t-1}) - (\beta-1)s(y_t, \vartheta_{t|t-1})s(y_t|\vartheta_{t|t-1})^\top\right)
\end{aligned}$$

and

$$\frac{\partial}{\partial \vartheta_{t|t-1}^\top}\|f_{t|t-1}\|_\beta^\beta = -\beta\|f_{t|t-1}\|_\beta^\beta \left(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1})\right)^\top.$$

Moreover,

$$\begin{aligned}
&\frac{\partial}{\partial \vartheta_{t|t-1}^\top} \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) \\
&\quad = \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t, \vartheta_{t|t-1}) - \int_{\mathcal{Y}} s(y|\vartheta_{t|t-1}) \frac{\partial}{\partial \vartheta_{t|t-1}^\top} \tilde{f}_{t|t-1}^\beta(y) \mu(\mathrm{d}y) \\
&\quad = \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t, \vartheta_{t|t-1}) - \beta\Bigg(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta}\left[s(y_t, \vartheta_{t|t-1})s(y_t, \vartheta_{t|t-1})^\top\right] \\
&\qquad - \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) \left(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1})\right)^\top\Bigg).
\end{aligned}$$

Hence,

$$
\begin{aligned}
&\frac{\partial}{\partial \vartheta_{t|t-1}^\top}\Big(\|f_{t|t-1}\|_\beta^\beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\Big)\\
&\quad= \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\frac{\partial}{\partial \vartheta_{t|t-1}^\top}\|f_{t|t-1}\|_\beta^\beta + \|f_{t|t-1}\|_\beta^\beta \frac{\partial}{\partial \vartheta_{t|t-1}^\top}\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\\
&\quad= -\beta\|f_{t|t-1}\|_\beta^\beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\left(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\right)^\top\\
&\qquad+ \|f_{t|t-1}\|_\beta^\beta \frac{\partial}{\partial \vartheta_{t|t-1}^\top}\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\\
&\quad= \|f_{t|t-1}\|_\beta^\beta \Bigg(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t,\vartheta_{t|t-1})\\
&\qquad- \beta\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\left(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\right)^\top\\
&\qquad- \beta\mathbb{E}_{\tilde{f}_{t|t-1}^\beta}\left[s(y_t,\vartheta_{t|t-1})s(y_t,\vartheta_{t|t-1})^\top\right]\\
&\qquad- \beta\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\left(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\right)^\top\Bigg).
\end{aligned}
$$

Then, combining terms,

$$
\begin{aligned}
&\nabla_\vartheta^2 \mathrm{PowS}_\beta(f_{t|t-1},y_t)\\
&\quad= \beta\frac{\partial}{\partial \vartheta_{t|t-1}^\top} f_{t|t-1}(y_t)^{\beta-1}s(y_t,\vartheta_{t|t-1}) - \beta^2\frac{\partial}{\partial \vartheta_{t|t-1}^\top}\Big(\|f_{t|t-1}\|_\beta^\beta \mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t,\vartheta_{t|t-1})\Big)\\
&\quad= \beta f_{t|t-1}(y_t)^{\beta-1}\left(H(y_t,\vartheta_{t|t-1}) - (\beta-1)s(y_t,\vartheta_{t|t-1})s(y_t|\vartheta_{t|t-1})^\top\right)\\
&\qquad- \beta^2\|f_{t|t-1}\|_\beta^\beta\left(\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} H(Y_t,\vartheta_{t|t-1}) - \beta\mathbb{E}_{\tilde{f}_{t|t-1}^\beta}\left[s(y_t,\vartheta_{t|t-1})s(y_t,\vartheta_{t|t-1})^\top\right]\right).
\end{aligned}
$$

B.2 Example 4

For the predictive density $f_{t|t-1}(y) = \frac{1}{\sigma_{t|t-1}}\varphi\left(\frac{y-\mu_{t|t-1}}{\sigma_{t|t-1}}\right)$, the L^β-norm is given by

$$
\begin{aligned}
\|f_{t|t-1}\|_\beta^\beta &= \int_{\mathbb{R}}\left(\frac{1}{\sigma_{t|t-1}}\varphi\left(\frac{y-\mu_{t|t-1}}{\sigma_{t|t-1}}\right)\right)^\beta \mathrm{d}y\\
&= \sigma_{t|t-1}^{1-\beta}\int_{\mathbb{R}}(2\pi)^{-\frac{\beta}{2}}\exp\left(-\frac{\beta}{2}z^2\right)\mathrm{d}z\\
&= \sigma_{t|t-1}^{1-\beta}(2\pi)^{\frac{1-\beta}{2}}\beta^{-\frac{1}{2}},
\end{aligned}
$$

and hence

$$\frac{f(y_t)^{\beta-1}}{\|f\|_\beta^{\beta-1}} = \frac{\sigma_{t|t-1}^{-(\beta-1)}(2\pi)^{-(\beta-1)/2}}{\sigma_{t|t-1}^{-\frac{(\beta-1)^2}{\beta}}(2\pi)^{-\frac{(\beta-1)^2}{2\beta}}\beta^{-\frac{\beta-1}{2\beta}}}\exp\left(-\tfrac{\beta-1}{2}z_t^2\right)$$
$$= \left(\frac{\sigma_{t|t-1}}{\sqrt{2\pi/\beta}}\right)^{-\frac{\beta-1}{\beta}}\sqrt{2\pi}\varphi\left(\sqrt{\beta-1}\,z_t\right),$$

where $z_t = (y_t - \mu_{t|t-1})/\sigma_{t|t-1}$. As desired, the ratio $\frac{f(y_t)^{\beta-1}}{\|f\|_\beta^{\beta-1}}$ tends to one if $\beta \downarrow 1$ since $\varphi(0) = 1/\sqrt{2\pi}$.

To obtain $\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1})$, we first realize that $\tilde{f}_{t|t-1}^\beta$ must be the density of a Gaussian random variable with mean $\mu_{t|t-1}$ and variance $\sigma_{t|t-1}^2/\beta$ since $f_{t|t-1}(y_t) \propto \exp\left(-\frac{(y_t-\mu_{t|t-1})^2}{2\sigma_{t|t-1}^2}\right)$. This immediately implies the first term in $\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1})$ to be zero. For the second term, we plug in the variance of $\tilde{Y}_t \sim \mathcal{N}(\mu_{t|t-1}, \sigma_{t|t-1}^2/\beta)$, which gives

$$\mathbb{E}_{\tilde{f}_{t|t-1}^\beta} s(Y_t, \vartheta_{t|t-1}) = \begin{pmatrix} 0 \\ \frac{1}{2} - \frac{1}{2\sigma_{t|t-1}^2}\mathbb{V}_{\tilde{f}_{t|t-1}^\beta}(\tilde{Y}_t) \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{\beta-1}{2\beta} \end{pmatrix}.$$

The implied gradient is

$$\nabla_\vartheta \mathrm{PsSphS}_\beta(f_{t|t-1}, y_t) = \left(\frac{\sigma_{t|t-1}}{\sqrt{2\pi/\beta}}\right)^{-\frac{\beta-1}{\beta}}\sqrt{2\pi}\varphi\left(\sqrt{\beta-1}z_t\right)\begin{pmatrix} -\frac{z_t}{\sigma_{t|t-1}} \\ \frac{1}{2}(1-z_t^2) - \frac{\beta-1}{2\beta} \end{pmatrix}.$$

B.3 Hyvärinen score

Using $\nabla_\vartheta \|s_y(y_t, \vartheta_{t|t-1})\|_2^2 = 2\big(\nabla_\vartheta s_y(y_t, \vartheta_{t|t-1})\big)^\top s_y(y_t, \vartheta_{t|t-1})$, we immediately arrive at

$$\nabla_\vartheta \mathrm{HS}(f_{t|t-1}, y_t) = \nabla_\vartheta \Delta_y \log f(y_t|\vartheta_{t|t-1}) + \frac{1}{2}\nabla_\vartheta \left\|\nabla_y \log f(y_t|\vartheta_{t|t-1})\right\|_2^2$$
$$= \Delta_y s(y_t, \vartheta_{t|t-1}) + \big(\nabla_\vartheta s_y(y_t, \vartheta_{t|t-1})\big)^\top s_y(y_t, \vartheta_{t|t-1}),$$

for the gradient. Here we recall the convention that the gradient of a vector-valued function $\bar{g}_m : \mathbb{R}^l \to \mathbb{R}^m$ is the $m \times l$ matrix $\nabla_x \bar{g}(x)$ with elements $[\nabla_x \bar{g}(x)]_{i,j} = \frac{\partial g_i(x)}{\partial x_j}$. In other words, $\nabla_\vartheta s_y(y_t, \vartheta_{t|t-1})$ is $d \times k$.

Similarly,

$$\begin{aligned}\nabla^2_\vartheta \mathrm{HS}(f_{t|t-1}, y_t) &= \Delta_y \nabla_\vartheta s(y_t, \vartheta) + \left(\nabla^2_\vartheta s_y(y_t, \vartheta_{t|t-1})\right)^\top s_y(y_t, \vartheta_{t|t-1}) \\ &\quad + \left(\nabla_\vartheta s_y(y_t, \vartheta_{t|t-1})\right) \times_3 \nabla_\vartheta s_y(y_t, \vartheta_{t|t-1}) \\ &= \Delta_y H(y_t, \vartheta_{t|t-1}) + \left(\nabla_y H(y_t, \vartheta_{t|t-1})\right) \times_3 s_y(y_t, \vartheta_{t|t-1}) \\ &\quad + \left(\nabla_\vartheta s_y(y_t, \vartheta_{t|t-1})\right)^\top \nabla_\vartheta s_y(y_t, \vartheta_{t|t-1}),\end{aligned}$$

where $\nabla_y H(y_t, \vartheta_{t|t-1})$ is a third order tensor of dimension $k \times k \times d$ and $\times_3$ is the 3-mode vector product of a tensor (Kolda and Bader 2009). Componentwise,

$$\left[\nabla_y H(y_t, \vartheta_{t|t-1}) \times_3 s_y(y_t, \vartheta_{t|t-1})\right]_{i_1, i_2} = \sum_{i_3=1}^{d} \frac{\partial H_{i_1, i_2}(y_t, \vartheta)}{\partial y_{t, i_3}} \left[s_y(y_t, \vartheta)\right]_{i_3},$$

for $i_1, i_2 = 1 \ldots, k$.

B.4 Kernel updates

For the gradient and Hessian of

$$S_\rho(\mathrm{F}_{t|t-1}, y_t) := \mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, y_t) - \frac{1}{2} \mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, X'_t) - \frac{1}{2} \rho(y_t, y_t),$$

we use the assumed interchangeability of differentiation and integration. Then

$$\begin{aligned}\nabla_\vartheta \mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, y_t) &= -\mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, y_t) s(X_t, \vartheta_{t|t-1})\right] \\ \nabla^2_\vartheta \mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, y_t) &= \mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, y_t) \left(s(X_t, \vartheta_{t|t-1}) s(X_t, \vartheta_{t|t-1})^\top - H(X_t, \vartheta_{t|t-1})\right)\right],\end{aligned}$$

and, similarly,

$$\begin{aligned}\nabla_\vartheta \mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, X'_t) &= -\mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, X'_t) \left(s(X_t, \vartheta_{t|t-1}) + s(X'_t, \vartheta_{t|t-1})\right)\right] \\ &= -2\mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, X'_t) s(X_t, \vartheta_{t|t-1})\right],\end{aligned}$$

and

$$
\begin{aligned}
&\nabla_\vartheta^2 \mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, X_t') \\
&= \mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, X_t') \big(\nabla_\vartheta s(X_t, \vartheta_{t|t-1}) + \nabla_\vartheta s(X_t', \vartheta_{t|t-1}) \big) \right] \\
&\quad + \mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, X_t') \big(s(X_t, \vartheta_{t|t-1}) + s(X_t', \vartheta_{t|t-1}) \big) \big(s(X_t, \vartheta_{t|t-1}) + s(X_t', \vartheta_{t|t-1}) \big)^\top \right] \\
&= 2\mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, X_t') \left(s(X_t, \vartheta_{t|t-1}) s(X_t, \vartheta_{t|t-1})^\top - H(X_t, \vartheta_{t|t-1}) \right) \right] \\
&\quad + 2\mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, X_t') s(X_t, \vartheta_{t|t-1}) s(X_t', \vartheta_{t|t-1})^\top \right].
\end{aligned}
$$

So,

$$
\begin{aligned}
\psi_{S_\rho}(y_t, \vartheta_{t|t-1}) &= \mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\left(\mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, X_t') - \rho(X_t, y_t) \right) s(X_t, \vartheta_{t|t-1}) \right], \\
H_{S_\rho}(y_t, \vartheta_{t|t-1}) &= \mathbb{E}_{\mathrm{F}_{t|t-1}} \Big[\left(\mathbb{E}_{\mathrm{F}_{t|t-1}} \rho(X_t, X_t') - \rho(X_t, y_t) \right) \big(H(X_t, \vartheta_{t|t-1}) \\
&\qquad - s(X_t, \vartheta_{t|t-1}) s(X_t, \vartheta_{t|t-1})^\top \big) \Big] \\
&\qquad - \mathbb{E}_{\mathrm{F}_{t|t-1}} \left[\rho(X_t, X_t') s(X_t, \vartheta_{t|t-1}) s(X_t', \vartheta_{t|t-1})^\top \right].
\end{aligned}
$$

B.5 Example 5

The gradient of

$$
\begin{aligned}
&\mathrm{CRPS}(\mathrm{F}_{t|t-1}, y_t) \\
&\quad = \sigma_{t|t-1} \left(\frac{y_t - \mu_{t|t-1}}{\sigma_{t|t-1}} \left(2\Phi \left(\frac{y_t - \mu_{t|t-1}}{\sigma_{t|t-1}} \right) - 1 \right) + 2\varphi \left(\frac{y_t - \mu_{t|t-1}}{\sigma_{t|t-1}} \right) - \frac{1}{\sqrt{\pi}} \right)
\end{aligned}
$$

with respect to the parameters reads

$$
\begin{aligned}
\nabla_\mu \mathrm{CRPS}(\mathrm{F}_{t|t-1}, y_t) &= \sigma_{t|t-1} \left((2\Phi(z_t) - 1) + 2 z_t \varphi(z_t) - z_t 2\varphi(z_t) \right) \cdot - \frac{1}{\sigma_{t|t-1}} \\
\nabla_\mu^2 \mathrm{CRPS}(\mathrm{F}_{t|t-1}, y_t) &= \frac{2}{\sigma_{t|t-1}} \varphi(z_t),
\end{aligned}
$$

where $z_t = (y_t - \mu_{t|t-1})/\sigma_{t|t-1}$, since $\nabla_\mu \text{CRPS}(\text{F}_{t|t-1}, y_t) = 1 - 2\Phi\left(z_t\right)$. Similarly,

$$\begin{aligned}
\nabla_\sigma \text{CRPS}(\text{F}_{t|t-1}, y_t) &= z_t\left(2\Phi\left(z_t\right) - 1\right) + 2\varphi\left(z_t\right) + \frac{1}{\sqrt{\pi}} - z_t\left(1 - 2\Phi(z_t)\right), \\
\nabla_\sigma^2 \text{CRPS}(\text{F}_{t|t-1}, y_t) &= \frac{2z_t^2}{\sigma_{t|t-1}^2}\varphi(z_t), \\
\nabla_\mu \nabla_\sigma \text{CRPS}(\text{F}_{t|t-1}, y_t) &= \frac{2z_t}{\sigma_{t|t-1}}\varphi(z_t),
\end{aligned}$$

since $\nabla_\sigma \text{CRPS}(\text{F}_{t|t-1}, y_t) = 2\varphi\left(z_t\right) - \frac{1}{\sqrt{\pi}}$, and hence $\nabla_\gamma \text{CRPS}(\text{F}_{t|t-1}, y_t) = \sigma_{t|t-1}\left(\varphi\left(z_t\right) - \frac{1}{2\sqrt{\pi}}\right)$, $\nabla_\mu \nabla_\gamma \text{CRPS}(\text{F}_{t|t-1}, y_t) = z_t\varphi(z_t)$, and $\nabla_\gamma^2 \text{CRPS}(\text{F}_{t|t-1}, y_t) = 2\sigma_{t|t-1}\left(z_t^2\varphi(z_t) + \varphi(z_t) - \frac{1}{2\sqrt{\pi}}\right)$.

C Derivations Section 3.5

C.1 Censored score updates

The gradient

$$\psi_{\text{LogS}_w^\flat}(y_t, \vartheta_{t|t-1}) = w(y_t)s(y_t, \vartheta_{t|t-1}) + \big(1 - w(y_t)\big)\mathbb{E}_{\text{F}_{1-w}^\sharp} s(Y_t, \vartheta_{t|t-1}),$$

follows directly from observing that

$$\begin{aligned}
\nabla_\vartheta \log \bar{F}_{\vartheta_{t|t-1}} &= \frac{1}{\bar{F}_{t|t-1}} \int_{\mathcal{Y}} \big(1 - w(y)\big)\nabla_\vartheta \log f(y|\vartheta_{t|t-1}) f(y|\vartheta_{t|t-1})\mu(\mathrm{d}y) \\
&= \int_{\mathcal{Y}} \nabla_\vartheta \log f(y|\vartheta_{t|t-1}) f_{1-w}^\sharp(y|\vartheta_{t|t-1})\mu(\mathrm{d}y) \\
&= -\mathbb{E}_{\text{F}_{1-w}^\sharp} \nabla_\vartheta \text{LogS}(f_{t|t-1}, Y_t),
\end{aligned}$$

since then

$$\begin{aligned}
\nabla_\vartheta \text{LogS}_w^\flat(f_{t|t-1}, y_t) &= -w(y_t)\nabla_\vartheta \log f(y_t|\vartheta_{t|t-1}) - \big(1 - w(y_t)\big)\nabla_\vartheta \log \bar{F}_{\vartheta_{t|t-1}} \\
&= w(y_t)\nabla_\vartheta \text{LogS}(f_{t|t-1}, y_t) - \big(1 - w(y_t)\big)\nabla_\vartheta \log \bar{F}_{\vartheta_{t|t-1}} \\
&= w(y_t)s(y_t, \vartheta_{t|t-1}) + \big(1 - w(y_t)\big)\mathbb{E}_{\text{F}_{1-w}^\sharp} s(Y_t, \vartheta_{t|t-1}).
\end{aligned}$$

C.2 Generalized censored logarithmic scoring rule

Consider a weight function w and nuisance density h. We first consider the gradient of the logarithmic score of the censored density $f^{\flat}_{w,\bar{h}}(y_t)$, for which we arrive at

$$
\begin{aligned}
&\nabla_\vartheta \log f^{\flat}_{w,\bar{h}}(y_t) \\
&\quad = \nabla_\vartheta \log \big(w(y_t) f_{t|t-1}(y_t) + \bar{F}_w \bar{h}(y_t)\big) \\
&\quad = \frac{w(y_t)\nabla_\vartheta f_{t|t-1}(y_t) + \bar{h}(y_t)\int (1-w)\nabla_\vartheta f_{t|t-1}\mathrm{d}\mu}{w(y_t) f_{t|t-1}(y_t) + \bar{F}_w \bar{h}(y_t)} \\
&\quad = \frac{w(y_t) f_{t|t-1}(y_t)\nabla_\vartheta \log f_{t|t-1}(y_t)}{w(y_t) f_{t|t-1}(y_t) + \bar{F}_w \bar{h}(y_t)} + \frac{\bar{h}(y_t)\int (1-w) f_{t|t-1}\nabla_\vartheta \log f_{t|t-1}\mathrm{d}\mu}{w(y_t) f_{t|t-1}(y_t) + \bar{F}_w \bar{h}(y_t)} \\
&\quad = \bar{w}(y_t) s(y_t, \vartheta_{t|t-1}) + \big(1-\bar{w}(y_t)\big)\mathbb{E}_{f^{\sharp}_{t|t-1}} s(Y_t, \vartheta_{t|t-1}),
\end{aligned}
$$

where

$$
\bar{w}(y_t) := \frac{w(y_t) f_{t|t-1}(y_t)}{w(y_t) f_{t|t-1}(y_t) + \bar{F}_w \bar{h}(y_t)} \in (0,1).
$$

Then,

$$
\begin{aligned}
&\nabla_\vartheta \mathrm{LogS}^{\flat}_{w,\bar{h}}(f_{t|t-1}, y_t) \\
&\quad = w(y_t)\nabla_\vartheta \log f^{\flat}_{w,\bar{h}}(y_t) + \big(1-w(y_t)\big)\mathbb{E}_{\bar{h}}\nabla_\vartheta \log f^{\flat}_{w,\bar{h}}(Q) \\
&\quad = w(y_t)\bar{w}(y_t) s(y_t, \vartheta_{t|t-1}) + \big(1-w(y_t)\big)\mathbb{E}_{\bar{h}}[\bar{w}(Y_t) s(Y_t, \vartheta_{t|t-1})] \\
&\qquad + \big(w(y_t)\big(1-\bar{w}(y_t)\big) + \big(1-w(y_t)\big)\mathbb{E}_{\bar{h}}[1-\bar{w}(Y_t)]\big)\,\mathbb{E}_{f^{\sharp}_{t|t-1}} s(Y_t, \vartheta_{t|t-1}) \\
&\quad = w_{\mathrm{a}}(y_t) s(y_t, \vartheta_{t|t-1}) + \big(1-w_{\mathrm{a}}(y_t) - w_{\mathrm{b}}(y_t)\big)\mathbb{E}_{f^{\sharp}_{t|t-1}} s(Y_t, \vartheta_{t|t-1}) \\
&\qquad + w_{\mathrm{b}}(y_t)\mathbb{E}_{\bar{h}^{\sharp}_{\bar{w}}}[s(Y_t, \vartheta_{t|t-1})],
\end{aligned}
$$

assuming $\int_{\mathcal{Y}} \bar{w}\bar{h}\mathrm{d}\mu > 0$, and defining

$$
w_{\mathrm{a}}(y_t) := w(y_t)\bar{w}(y_t), \quad w_{\mathrm{b}}(y_t) := \big(1-\bar{w}(y_t)\big)\mathbb{E}_{\bar{h}}\bar{w}(Y_t).
$$

Now, if $\bar{h} = \delta_*$ on $\mathcal{Y}^*$, with $w(*) = 0$, then $\bar{w}(y_t) = 1$ for all $y \in \mathcal{Y}$ since $\delta_*(y_t) = 0$ for all $y \in \mathcal{Y}$, and $\bar{w}(*) = 0$. Hence, $w_{\mathrm{a}}(y_t) = w(y_t)$ on $\mathcal{Y}^*$. Additionally, $\mathbb{E}_h \bar{w}(Y_t) = \bar{w}(*) = 0$

implies $w_{\mathrm{b}}(y_t) = 0$ for all $y \in \mathcal{Y}^*$. Therefore,

$$\nabla_\vartheta \mathrm{LogS}^\flat_{w,\delta_*}(f_{t|t-1}, y_t) = w(y_t)s(y_t, \vartheta_{t|t-1}) + \big(1 - w(y_t)\big)\mathbb{E}_{f^\sharp_{t|t-1}} s(Y_t, \vartheta_{t|t-1}),$$

coinciding with $\psi_{\mathrm{LogS}^\flat_w}(y_t, \vartheta_{t|t-1})$ in Equation (8).

Moreover, if the support of $\bar{h}$ is a subset of $R^c_w \subseteq \mathcal{Y}$, then $\int_{\mathcal{Y}} \bar{w}(y)\bar{h}(y)\mu(\mathrm{d}y) = 0$ for all $y \in \mathcal{Y}$. Consequently, $w_{\mathrm{b}}(y_t) = 0$ for all $y_t \in \mathcal{Y}$ and the third term vanishes, so that the condition $\int_{\mathcal{Y}} \bar{w}\bar{h}\mathrm{d}\mu > 0$ becomes irrelevant. In this case, the dependence on $\bar{h}$ drops out, which agrees with the $\bar{h}$-independent form of $\mathrm{LogS}^\flat_{w,\bar{h}}$ in Table 1 of De Punder et al. (2025), where the same support restriction on $\bar{h}$ is imposed. If $\bar{h}$ and f are both μ-densities, strict local propriety of $\mathrm{LogS}^\flat_{w,\bar{h}}$ requires $\mu(\{y \in \mathcal{Y} : \bar{h}(y) > 0, w(y)f(y) = 0\}) > 0$, for which the latter restriction on $\bar{h}$ is sufficient, but generally not necessary.

C.3 Generalized censored Hyvärinen score

The censored Hyvärinen score reads

$$\begin{aligned}
&\mathrm{HS}^\flat_{w,\bar{h}}(f_{t|t-1}, y_t) \\
&\quad = w(y_t)\mathrm{HS}(f^\flat_{w,\bar{h}}, y_t) + \big(1 - w(y_t)\big) \int_{-\infty}^{\infty} \mathrm{HS}(f^\flat_{w,\bar{h}}, q)\varphi(q)\mathrm{d}q \\
&\quad = w(y_t) \left(\Delta_y \log f^\flat_{w,\bar{h}}(y_t) + \frac{1}{2} \left\| \nabla_y \log f^\flat_{w,\bar{h}}(y_t) \right\|_2^2 \right) \\
&\qquad + \big(1 - w(y_t)\big) \int_{-\infty}^{\infty} \left(\Delta_q \log f^\flat_{w,\bar{h}}(q) + \frac{1}{2} \left\| \nabla_q \log f^\flat_{w,\bar{h}}(q) \right\|_2^2 \right) \varphi(q)\mathrm{d}q.
\end{aligned}$$

Therefore, the gradient becomes

$$\begin{aligned}
&\nabla_\vartheta \mathrm{HS}^\flat_{w,\bar{h}}(f_{t|t-1}, y_t) \\
&\quad = w(y_t) \left(\Delta_y \nabla_\vartheta \log f^\flat_{w,\bar{h}}(y_t) + \frac{1}{2} \nabla_\vartheta \left\| \nabla_y \log f^\flat_{w,\bar{h}}(y_t) \right\|_2^2 \right) \\
&\qquad + \big(1 - w(y_t)\big) \int_{-\infty}^{\infty} \left(\Delta_q \nabla_\vartheta \log f^\flat_{w,\bar{h}}(q) + \frac{1}{2} \nabla_\vartheta \left\| \nabla_q \log f^\flat_{w,\bar{h}}(q) \right\|_2^2 \right) \varphi(q)\mathrm{d}q.
\end{aligned}$$

From Appendix C.2, we recall the gradient of the generalized censored logarithmic scoring rule for general h, given by

$$\nabla_\vartheta \log f^\flat_{w,\bar{h}}(y_t) = \bar{w}(y_t) s(y_t, \vartheta_{t|t-1}) + \big(1 - \bar{w}(y_t)\big) \mathbb{E}_{f^\sharp_{t|t-1}} s(Y_t, \vartheta_{t|t-1}),$$

where

$$\bar{w}(y_t) = \frac{w(y_t) f_{t|t-1}(y_t)}{w(y_t) f_{t|t-1}(y_t) + \bar{F}_w \bar{h}(y_t)} \in (0,1),$$

with gradient

$$\nabla_y \bar{w}(y_t) = \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big)\Big(s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t) - s^{\bar{h}}_y(y_t)\Big),$$

where $s^{\bar{h}}_y(y_t) = \nabla_y \log \bar{h}(y_t)$ and $s^w_y(y_t) = \nabla_y \log w(y_t)$. By differentiating $\nabla_y \bar{w}(y_t)$, we obtain the Laplacian

$$\begin{aligned}\Delta_y \bar{w}(y_t) = {} & \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big)\left(\Delta_y \log w(y_t) + \Delta_y \log f_{t|t-1}(y_t) - \Delta_y \log \bar{h}(y_t)\right) \\ & + \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big)\big(1 - 2\bar{w}(y_t)\big)\left\| s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t) - s^{\bar{h}}_y(y_t)\right\|_2^2.\end{aligned}$$

Following the steps in Appendix B.3 yields

$$\begin{aligned}\Delta_y \nabla_\vartheta \log f^\flat_{w,\bar{h}}(y_t) &= \Delta_y \left\{\bar{w}(y_t)\Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{f^\sharp_{t|t-1}} s(Y_t, \vartheta_{t|t-1})\Big)\right\} \\ &= \bar{w}(y_t) \Delta_y s(y_t, \vartheta_{t|t-1}) + 2\big(\nabla_y \bar{w}(y_t)\big) \times_3 \nabla_y s(y_t, \vartheta_{t|t-1}) \\ &\quad + \Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{f^\sharp_{t|t-1}} s(Y_t, \vartheta_{t|t-1})\Big) \Delta_y \bar{w}(y_t),\end{aligned}$$

and

$$\frac{1}{2} \nabla_\vartheta \left\| \nabla_y \log f^\flat_{w,\bar{h}}(y_t) \right\|_2^2 = \Big(\nabla_y \log f^\flat_{w,\bar{h}}(y_t)\Big)^\top \nabla_\vartheta \nabla_y \log f^\flat_{w,\bar{h}}(y_t),$$

where

$$\begin{aligned}\nabla_y \log f^\flat_{w,\bar{h}}(y_t) &= \frac{\big(\nabla_y w(y_t)\big) f_{t|t-1}(y_t) + w(y_t) \nabla_y f_{t|t-1}(y_t) + \bar{F}_w \nabla_y \bar{h}(y_t)}{w(y_t) f_{t|t-1}(y_t) + \bar{F}_w \bar{h}(y_t)} \\ &= \bar{w}(y_t) s_y(y_t, \vartheta_{t|t-1}) + \big(1 - \bar{w}(y_t)\big) s^{\bar{h}}_y(y_t) + \bar{w}(y_t) s^w_y(y_t),\end{aligned}$$

and

$$\begin{aligned}\Big(\nabla_\vartheta \nabla_y \log f^\flat_{w,\bar{h}}(y_t)\Big)^\top &= \nabla_y \nabla_\vartheta \log f^\flat_{w,\bar{h}}(y_t) \\ &= d^\sharp_s(y_t, \vartheta_{t|t-1})\big(\nabla_y \bar{w}(y_t)\big)^\top + \bar{w}(y_t)\nabla_y s(y_t, \vartheta_{t|t-1}),\end{aligned}$$

where $d^\sharp_s(y_t, \vartheta_{t|t-1}) := s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{f^\sharp_{t|t-1}} s(Y_t, \vartheta_{t|t-1})$. Consequently,

$$\begin{aligned}&\frac{1}{2}\nabla_\vartheta \left\|\nabla_y \log f^\flat_{w,\bar{h}}(y_t)\right\|_2^2 \\ &= \Big(\nabla_\vartheta \nabla_y \log f^\flat_{w,\bar{h}}(y_t)\Big)^\top \nabla_y \log f^\flat_{w,\bar{h}}(y_t) \\ &= \Big(d^\sharp_s(y_t, \vartheta_{t|t-1})\big(\nabla_y \bar{w}(y_t)\big)^\top + \nabla_y s(y_t, \vartheta_{t|t-1})\bar{w}(y_t)\Big)\Big(\nabla_y \log f^\flat_{w,\bar{h}}(y_t)\Big) \\ &= c_1(y_t, \vartheta_{t|t-1}) d^\sharp_s(y_t, \vartheta_{t|t-1}) + \nabla_y s(y_t, \vartheta_{t|t-1}) c_2(y_t, \vartheta_{t|t-1}),\end{aligned}$$

where

$$\begin{aligned}c_1(y_t, \vartheta_{t|t-1}) &:= \bar{w}(y_t) s_y(y_t, \vartheta_{t|t-1})^\top \nabla_y \bar{w}(y_t) + \big(1 - \bar{w}(y_t)\big) s^{\bar{h}}_y(y_t)^\top \nabla_y \bar{w}(y_t) \\ &\quad + \bar{w}(y_t) s^w_y(y_t)^\top \nabla_y \bar{w}(y_t) \\ &= \Big(\bar{w}(y_t)\big(s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t)\big) + \big(1 - \bar{w}(y_t)\big) s^{\bar{h}}_y(y_t)\Big)^\top \nabla_y \bar{w}(y_t) \\ c_2(y_t, \vartheta_{t|t-1}) &:= \bar{w}(y_t)^2 s_y(y_t, \vartheta_{t|t-1}) + \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big) s^{\bar{h}}_y(y_t) + \bar{w}(y_t)^2 s^w_y(y_t) \\ &= \bar{w}(y_t)\Big(\bar{w}(y_t)\big(s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t)\big) + \big(1 - \bar{w}(y_t)\big) s^{\bar{h}}_y(y_t)\Big).\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\nabla_\vartheta \mathrm{HS}(f^\flat_{w,\bar{h}}, y_t) \\
&\quad = \bar{w}(y_t)\Delta_y s(y_t, \vartheta_{t|t-1}) + 2\big(\nabla_y \bar{w}(y_t)\big) \times_3 \nabla_y s(y_t, \vartheta_{t|t-1}) \\
&\qquad + \Big(\big(\Delta_y \bar{w}(y_t)\big) + c_1(y_t, \vartheta_{t|t-1})\Big) d^\sharp_s(y_t, \vartheta_{t|t-1}) \\
&\qquad + \nabla_y s(y_t, \vartheta_{t|t-1}) c_2(y_t, \vartheta_{t|t-1}) \\
&\quad = \bar{w}(y_t)\left(\Delta_y s(y_t, \vartheta_{t|t-1}) + \bar{w}(y_t)\nabla_y s(y_t, \vartheta_{t|t-1})\left(s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t)\right)\right) \\
&\qquad + \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big) \\
&\qquad + \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big) c_4(y_t, \vartheta_{t|t-1})\Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{f^\sharp_{t|t-1}} s(Y_t, \vartheta_{t|t-1})\Big) \\
&\quad = \bar{w}(y_t)\left(\Delta_y s(y_t, \vartheta_{t|t-1}) + \bar{w}(y_t)\nabla_y s(y_t, \vartheta_{t|t-1})\left(s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t)\right)\right) \\
&\qquad + \bar{w}(y_t)\big(1 - \bar{w}(y_t)\big) c_3(y_t, \vartheta_{t|t-1}),
\end{aligned}
$$

where $s^{f+w-h}_y(y_t, \vartheta_{t|t-1}) := s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t) - s^{\bar{h}}_y(y_t)$, and

$$
\begin{aligned}
c_3(y_t, \vartheta_{t|t-1}) := &\Big(\nabla_y s(y_t, \vartheta_{t|t-1}) s^{\bar{h}}_y(y_t) + 2 s^{f+w-h}_y(y_t, \vartheta_{t|t-1}) \times_3 \nabla_y s(y_t, \vartheta_{t|t-1})\Big) \\
&+ c_4(y_t, \vartheta_{t|t-1})\Big(s(y_t, \vartheta_{t|t-1}) - \mathbb{E}_{f^\sharp_{t|t-1}} s(Y_t, \vartheta_{t|t-1})\Big),
\end{aligned}
$$

where

$$
\begin{aligned}
&c_4(y_t, \vartheta_{t|t-1}) \\
&\quad = \Delta_y \log f_{t|t-1}(y_t) + \Delta_y \log w(y_t) - \Delta_y \log \bar{h}(y_t) \\
&\qquad + \big(1 - 2\bar{w}(y_t)\big) \left\| s^{f+w-h}_y(y_t, \vartheta_{t|t-1}) \right\|_2^2 \\
&\qquad + \Big(\bar{w}(y_t)\big(s_y(y_t, \vartheta_{t|t-1}) + s^w_y(y_t)\big) + \big(1 - \bar{w}(y_t)\big) s^{\bar{h}}_y(y_t)\Big)^\top s^{f+w-h}_y(y_t, \vartheta_{t|t-1}).
\end{aligned}
$$

Note that $c_4(y_t, \vartheta_{t|t-1})$ is a function of $s_y(y_t, \vartheta_{t|t-1})$, $\nabla_y s(y_t, \vartheta_{t|t-1})$, $s^w_y(y_t)$, $s^{\bar{h}}_y(y_t)$, $\bar{w}(y_t)$ only, i.e.,

$$
c_3(y_t, \vartheta_{t|t-1}) \equiv \Xi\big(s_y(y_t, \vartheta_{t|t-1}), \nabla_y s(y_t, \vartheta_{t|t-1}), s^w_y(y_t), s^{\bar{h}}_y(y_t), \bar{w}(y_t)\big).
$$

The gradient of the Hyvärinen score then follows by substituting the expression for $\nabla_\vartheta \mathrm{HS}(f^\flat_{t|t-1}, y_t)$ into

$$\nabla_\vartheta \mathrm{HS}^\flat_{w,\bar{h}}(f_{t|t-1}, y_t) = w(y_t)\nabla_\vartheta \mathrm{HS}(f^\flat_{t|t-1}, y_t) + \big(1 - w(y_t)\big) \int_{-\infty}^{\infty} \nabla_\vartheta \mathrm{HS}(f^\flat_{t|t-1}, q)\varphi(q)\mathrm{d}q.$$

D Set-valued functionals

Following Section 2.1, we temporarily suppress time indices and let $\mathcal{Y} \subseteq \mathbb{R}^l$. Consider a (potentially) set valued functional $\tilde{u} : \mathcal{P} \twoheadrightarrow \mathcal{Y}$, and recall that strict consistency of a scoring function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with respect to $(\mathcal{P}, \tilde{u})$ was already defined in Section 2.1. Strict consistency requires that

$$\tilde{u}(\mathrm{P}) = \arg\min_{x \in \mathcal{Y}} \mathbb{E}_\mathrm{P} \ell(x, Y), \quad \forall \mathrm{P} \in \mathcal{P},$$

that is, $\mathbb{E}_\mathrm{P}\ell(x_\mathrm{P}, Y) \leq \mathbb{E}_\mathrm{P}\ell(x, Y)$, for all $x_\mathrm{P} \in \tilde{u}(\mathrm{P})$ and $x \in \mathcal{Y}$, and $\mathbb{E}_\mathrm{P}\ell(x_\mathrm{P}, Y) = \mathbb{E}_\mathrm{P}\ell(x, Y)$ implies $x \in \tilde{u}(\mathrm{P})$; see Definition 1 in Gneiting (2011).

A strictly consistent scoring function ℓ for $\tilde{u} : \mathcal{P} \twoheadrightarrow \mathcal{Y}$ induces a family of scoring rules $\mathcal{S}_\ell(\mathrm{F}, y) := \{S_\ell(\mathrm{F}, y; x_\mathrm{F}) : x_\mathrm{F} \in \tilde{u}(\mathrm{F})\}$, where $S_\ell(\mathrm{F}, y; x_\mathrm{F}) := \ell(x_\mathrm{F}, y)$, for all $x_\mathrm{F} \in \tilde{u}(\mathrm{F})$. Ordinary scoring rules arise as degenerate families consisting of a single member.

As generalizations of Definitions 1 and 2, we first formalize the notion of strict local propriety and local divergences for set-valued functionals.

Definition 8. *Let $\tilde{u} : \mathcal{P} \twoheadrightarrow \mathcal{Y}$ be a functional. The family of scoring rules $\mathcal{S}_\ell$ is called* strictly locally proper with respect to $(\mathcal{P}, \tilde{u})$ *if (i) $S_\ell(\mathrm{P}, Y; x_\mathrm{P}) = S_\ell(\mathrm{F}, y; x_\mathrm{F})$ whenever $\tilde{u}(\mathrm{P}) = \tilde{u}(\mathrm{F})$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$; and (ii) $\sup_{x_\mathrm{P} \in \tilde{u}(\mathrm{P})} \mathbb{E}_\mathrm{P} S_\ell(\mathrm{P}, Y; x_\mathrm{P}) \leq \inf_{x_\mathrm{F} \in \tilde{u}(\mathrm{F})} \mathbb{E}_\mathrm{P} S_\ell(\mathrm{F}, Y; x_\mathrm{F})$, with equality if and only if $\tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F}) \neq \emptyset$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$.*

Definition 9. *Let $\tilde{u} : \mathcal{P} \twoheadrightarrow \mathcal{Y}$ be a functional. A map $\mathbb{D} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ is called a* local divergence with respect to $\tilde{u}$ *if (i) $\tilde{u}(\mathrm{F}) = \tilde{u}(\mathrm{G})$ implies $\mathbb{D}(\mathrm{P}\|\mathrm{F}) = \mathbb{D}(\mathrm{P}\|\mathrm{G})$, $\forall \mathrm{P}, \mathrm{F}, \mathrm{G} \in \mathcal{P}$, (ii) $\mathbb{D}(\mathrm{P}\|\mathrm{F}) \geq 0$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$, and (iii) $\mathbb{D}(\mathrm{P}\|\mathrm{F}) = 0$ if and only if $\tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F}) \neq \emptyset$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$.*

Definition 9 extends to functionals $\widetilde{\mathcal{T}} : \mathcal{P} \to \mathcal{V}$, although this generalization is not pursued here. The next result connects the strict local propriety of scoring rule families to the strict consistency of the scoring functions by which they are induced. Proposition 4 then established that any strictly locally proper scoring rule family induces a local divergence.

Proposition 3. *Consider a family of scoring rules $\mathcal{S}_\ell : \mathcal{P} \times \mathcal{Y} \twoheadrightarrow \bar{\mathbb{R}}$. Assume that $\forall y \in \mathcal{Y}$ $\exists \mathrm{F}_y \in \mathcal{P}$ with $\tilde{u}(\mathrm{F}_y) = \{y\}$. Then $\mathcal{S}_\ell$ is strictly locally proper with respect to $(\mathcal{P}, \tilde{u})$ if and only if ℓ is strictly consistent with respect to $(\mathcal{P}, \tilde{u})$.*

Proposition 4. *Assume a family of scoring rules $\mathcal{S}_\ell : \mathcal{P} \times \mathcal{Y} \twoheadrightarrow \bar{\mathbb{R}}$ is strictly locally proper with respect to $\tilde{u} : \mathcal{P} \twoheadrightarrow \mathcal{Y}$. Then $\mathbb{D}_{S_\ell}(\mathrm{P}\|\mathrm{F}; x_\mathrm{P}) := \inf_{x \in \tilde{u}(\mathrm{F})}\{\mathbb{E}_\mathrm{P} S_\ell(\mathrm{F}, Y; x) - \mathbb{E}_\mathrm{P} S_\ell(\mathrm{P}, Y; x_\mathrm{P})\}$ is a local divergence with respect to $\tilde{u}$, for all $x_\mathrm{P} \in \tilde{u}(\mathrm{P})$.*

From the perspective of Bayes acts, the action space $\mathcal{A} := \{(\mathrm{F}, x_\mathrm{F}) : \mathrm{F} \in \mathcal{P}, x_\mathrm{F} \in \tilde{u}(\mathrm{F})\}$ is reduced to $\mathcal{A} = \mathcal{P}$, by selecting the best member via $\inf_{x \in \tilde{u}(\mathrm{F})} \mathbb{E}_\mathrm{P} S_\ell(\mathrm{F}, Y; x)$. In this case, $B_{S_\ell}(\mathrm{P}) = \{\mathrm{F} \in \mathcal{P} : \tilde{u}(\mathrm{F}) \cap \tilde{u}(\mathrm{P}) \neq \emptyset\}$. If $v(\mathrm{F})$ is a singleton $\forall \mathrm{F} \in \mathcal{P}$, the Bayes act further reduces to $B_{S_\ell}(\mathrm{P}) = \{\mathrm{F} \in \mathcal{P} : \tilde{u}(\mathrm{F}) = \tilde{u}(\mathrm{P})\}$. Since S_ℓ depends on F only through $\tilde{u}(\mathrm{F})$, the Bayes act is unique only up to $\tilde{u}$.

We now extend Corollary 1 to members of families of scoring rules $\mathcal{S}_\ell$ that are strictly locally proper with respect to $(\mathcal{P}, \tilde{u})$, for which we reintroduce time indices. A member $S_\ell(\mathrm{F}_{t|t-1}, y_t; x_\mathrm{F}) \in \mathcal{S}_\ell$ for some specific $x_\mathrm{F} \in \tilde{u}(\mathrm{F}_{t|t-1})$ is implictly selected through a measurable *selector* with $\zeta(R) \in R$ for every nonempty $R \subseteq \mathcal{Y}$, and such that $\mathrm{F}_{t|t-1} \mapsto \zeta \circ \tilde{u}(\mathrm{F}_{t|t-1})$, is $\mathcal{F}_{t-1}$-measurable for all $\mathrm{F}_{t|t-1} \in \mathcal{P}$. The composition $\zeta \circ \tilde{u}$ is not an ordinary functional

in the sense of a deterministic mapping from $\mathcal{P}$ onto (subsets of) $\mathcal{Y}$, but a selection device applied to the correspondence $\mathrm{F}_{t|t-1} \mapsto \tilde{u}(\mathrm{F}_{t|t-1})$. Each selector induces a ζ-selected scoring rule $S_\ell^\zeta : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$.

Corollary 4. *Consider a functional $\tilde{u} : \mathcal{P} \twoheadrightarrow \mathcal{Y}$. Let $\mathcal{S}_\ell$ be a scoring rule family that is strictly locally proper with respect to $(\mathcal{P}, \tilde{u})$, with ζ^*-selected scoring rules $S_\ell^{\zeta^*} : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ implied by the selector ζ^* such that $\zeta^* \circ \tilde{u} \in \arg\min_{x \in \tilde{u}(\mathrm{F}_{t|t-1})} \mathbb{E}_{\mathrm{P}_t} S_\ell(\mathrm{F}_{t|t-1}, Y_t; x)$, for all $\mathrm{F}_{t|t-1} \in \mathcal{P}$. Assume $\mathbb{E}_{\mathrm{P}_t} \psi_{S_\ell^{\zeta^*}}(Y_t, \vartheta_{t|t-1})$ to be non-zero and let Assumption 1 and 2 hold. Then $\phi_{S_\ell^{\zeta^*}}(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1} \psi_{S_\ell^{\zeta^*}}(y_t, \vartheta_{t|t-1})$ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}(\cdot\|\cdot; \zeta^* \circ \tilde{u}(\cdot)), \tilde{u})$.*

In Corollary 4, the role of the measurable selector ζ^* is purely technical but unavoidable. When $\tilde{u}(\mathrm{F}_{t|t-1})$ is not a singleton, as with the β quantile of a distribution function that is flat at level β, no scoring rule can strictly locally proper for all $x \in \tilde{u}(\mathrm{F}_{t|t-1})$. This corresponds to the second case in Footnote 2 of Dimitriadis et al. (2024), where the identification remains strict for the set valued quantile $q_\beta(\mathrm{F}_{t|t-1})$ but not for any single valued representative such as $\mathrm{VaR}_\beta(\mathrm{F}_{t|t-1}) = \inf q_\beta(\mathrm{F}_{t|t-1})$. The selector ζ^* chooses a measurable representative $\zeta^*(\mathrm{F}_{t|t-1})$, inducing a single-valued scoring rule $S_\ell^{\zeta^*}(\mathrm{F}_{t|t-1}, y_t) := S_\ell(\mathrm{F}_{t|t-1}, y_t)$. Since $\mathcal{S}_\ell$ is strictly locally proper with respect to $(\mathcal{P}, \tilde{u})$, the induced $S_\ell^{\zeta^*}$ inherits this property by the same argument as in Proposition 4.

E Additional examples

E.1 Unweighted Hyvärinen score

We supplement the discussion of the Hyvärinen (2005) score in Section 3.4, with two examples. As first example, we revisit the univariate Gaussian location-scale model also

considered in Examples 3, 4 and 5, with mean $\mu_{t|t-1}$ and variance $\sigma_{t|t-1} = \exp(\gamma_{t|t-1})$. As shown by the gradient and Hessian in Example E.1.1, the second-order locality of the Hyvärinen (2005) induces a scaling proportional to $1/\sigma^2_{t|t-1}$, hence dampening updates if the variance is larger than one. This scaling, however, does not address the unboundedness of the update. Consequently, the robustness properties of the Hyvärinen score are closer to those of the logarithmic scoring rule than to the bounded updates obtained for the PsSphS and CRPS in Examples 4 and and 5, respectively. This observation aligns with the intuition that the Hyvärinen score remains local, though with respect to a neighborhood of the realization rather than the realization itself.

Example E.1.1. *For the Gaussian univariate location-scale model with mean $\mu_{t|t-1}$ and variance $\sigma^2_{t|t-1} = \exp(\gamma_{t|t-1})$, we arrive at*

$$\nabla_\vartheta \mathrm{HS} = \frac{1}{\sigma^2_{t|t-1}} \begin{pmatrix} s_\mu(y_t, \vartheta_{t|t-1}) \\ 2s_\gamma(y_t, \vartheta_{t|t-1}) \end{pmatrix}$$

$$\nabla^2_\vartheta \mathrm{HS} = \frac{1}{\sigma^2_{t|t-1}} \begin{pmatrix} H_{\mu\mu}(y_t, \vartheta_{t|t-1}) & 2H_{\mu\gamma}(y_t, \vartheta_{t|t-1}) \\ 2H_{\mu\gamma}(y_t, \vartheta_{t|t-1}) & 4H_{\gamma\gamma}(y_t, \vartheta_{t|t-1}) - 1 \end{pmatrix}.$$

The higher-order locality of the Hyvärinen score leads to a rescaling of the update and Hessian by the inverse variance, which reduces sensitivity to extreme observations, although it does not eliminate the unboundedness of the Hessian.

A similar conclusion applies to the Gaussian multivariate local level model discussed in the following example.

Example E.1.2. *For the Gaussian multivariate local level model with mean $\mu_{t|t-1} \in \mathbb{R}^k$ and fixed variance $\Sigma \succ O_k$. Since $\Delta_y \log f_{t|t-1}(y_t) = -\operatorname{tr}(\Sigma^{-1})$ and $\nabla_y \log f_{t|t-1}(y_t) =$*

$-\Sigma^{-1}(y_t - \mu_{t|t-1})$, *it follows that*

$$\mathrm{HS}\big(f_{t|t-1}, y_t\big) = -\operatorname{tr}(\Sigma^{-1}) + \tfrac{1}{2}(y_t - \mu_{t|t-1})^\top \Sigma^{-2}(y_t - \mu_{t|t-1}).$$

Hence, $\psi_{\mathrm{HS}}(y_t, \mu_{t|t-1}) = -\Sigma^{-2}(y_t - \mu_{t|t-1})$ *and* $H_{\mathrm{HS}}(y_t, \mu_{t|t-1}) = \Sigma^{-1} H(y_t, \mu_{t|t-1})$. *Relative to the logarithmic scoring rule, deviations of* y_t *from* $\mu_{t|t-1}$ *are therefore penalized in units of squared variance rather than variance under the the Hyvärinen score.*

E.2 Censored likelihood score

Example E.2.1, revisits the unbounded update in Example 3, applying the center indicator function of Section 4.3.3 to obtain bounded updates. Updates outside the compact interval $[r_1, r_2]$ only enter the update through $\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} s(Y_t, \vartheta_{t|t-1})$, which is bounded in the observation space. Boundedness on compact subsets of the parameter space then suffices, making the term $1/\sigma_{t|t-1}$ unproblematic for obtaining a PRADA update (see Theorem 3.2).

Instead of hard thresholds, one may also use smooth weight functions such as the Gaussian kernel $y \mapsto \phi(y)/\phi(0)$. The gradients and Hessians are then robustified in the same spirit as the PsSphS and CRPS implied updates introduced in Examples 4.4 and 4.5.

Example E.2.1. *Reconsider the Gaussian location-scale model introduced in Example 3. For the weight function* $w_1(y) = \mathbb{1}_{[r_1, r_2]}(y)$, *the gradient becomes*

$$\nabla_\vartheta \mathrm{LogS}^\flat_{w_1} = w_1(y_t) \begin{pmatrix} -\frac{z_t}{\sigma_{t|t-1}} \\ \frac{1}{2}(1 - z_t^2) \end{pmatrix} + \big(1 - w_1(y_t)\big) \begin{pmatrix} \frac{1}{\sigma_{t|t-1}} \frac{\varphi(\bar{z}_1) - \varphi(\bar{z}_2)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} \\ \frac{1}{2} \frac{\bar{z}_1 \varphi(\bar{z}_1) - \bar{z}_2 \varphi(\bar{z}_2)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} \end{pmatrix}$$

$$\nabla^2_{\vartheta}\mathrm{LogS}^{\flat}_{w_1} = w_1(y_t)\begin{pmatrix} \frac{1}{\sigma^2_{t|t-1}} & \frac{z_t}{\sigma_{t|t-1}} \\ \frac{z_t}{\sigma_{t|t-1}} & \frac{1}{2}z_t^2 \end{pmatrix} + \big(1-w_1(y_t)\big)\begin{pmatrix} \frac{1}{\sigma^2_{t|t-1}} & \frac{1}{\sigma_{t|t-1}}\frac{\varphi(\bar{z}_2)-\varphi(\bar{z}_1)}{\Phi(\bar{z}_1)+1-\Phi(\bar{z}_2)} \\ \frac{1}{\sigma_{t|t-1}}\frac{\varphi(\bar{z}_2)-\varphi(\bar{z}_1)}{\Phi(\bar{z}_1)+1-\Phi(\bar{z}_2)} & \frac{1}{2}\left(1+\frac{\bar{z}_2\varphi(\bar{z}_2)-\bar{z}_1\varphi(\bar{z}_1)}{\Phi(\bar{z}_1)+1-\Phi(\bar{z}_2)}\right) \end{pmatrix},$$

where $\bar{z}_j := \frac{r_j-\mu_{t|t-1}}{\sigma_{t|t-1}}$ *for* $j \in \{1,2\}$.

E.3 Smooth approximation indicator weight function

The following example illustrates the role of smooth indicator approximations in ensuring the required twice continuous differentiability of the density by the Hyvärinen (2005) score. In the limit, the smooth approximation converges to the discontinuous indicator function, while maintaining differentiability for any finite smooth parameter.

Example E.3.1. *Consider the weight function* $w_2(y;a) = \Psi\left(\frac{y-r_1}{a}\right) - \Psi\left(\frac{y-r_2}{a}\right)$, *where* $\Psi(x) = \int_{-\infty}^{x} b(s)\mathrm{d}s / \int_{-\infty}^{\infty} b(s)\mathrm{d}s$, *with* $b(s) = \exp(-1/(1-s^2))\mathbb{1}_{\{|s|<1\}}$ *and* $a > 0$. *The weight function is a smoothed analogue of* w_1, *satisfying* $w_2(y;a) = 1$ *for* $y \in (r_1, r_2)$ *and* $w_2(y;a) = 0$ *for* $y \le r_1 - a$ *or* $y \ge r_2 + a$. *Combined with the nuisance density* $\bar{h}(y) = \varphi(y)$, *the generalized censored density* $f^{\flat}_{w_2,\varphi}(y) = w(y)f(y) + \bar{F}_w\varphi(y)$ *is twice continuously differentiable in* y. *Moreover, Assumption 1 in De Punder et al. (2025) is satisfied for this combination of* w_2 *and* $\bar{h}_2$, *while it fails in the limit* $\lim_{a\downarrow 0} w_2(y;\alpha) = w_1(y)$. *In practice, one can choose* $a > 0$ *sufficiently small so that* $w_2(\cdot;a)$ *approximates* w_1, *mitigating the effect through* Ξ.

F Pseudo-target and pseudo-truth

We display results in a static context. Consider a scoring rule $S : \mathcal{P}_\Theta \times \mathcal{Y} \to \bar{\mathbb{R}}$, where

$$\mathcal{P}_\Theta \coloneqq \{\mathrm{P}_\vartheta : \vartheta \in \Theta\},$$

denotes a parametric family of distributions. We first consider correct specification, where the true distribution can be written as P_{ϑ_0} for some $\vartheta_0 \in \Theta$. We then turn to the potentially distributionally misspecified case, where $\mathrm{P} \in \mathcal{P}$ but not necessarily $\mathrm{P} \in \mathcal{P}_\Theta$.

F.1 Correct specification

Suppose that the true distribution belongs to the postulated model class, that is, $\mathrm{P} = \mathrm{P}_{\vartheta_0}$ for some $\vartheta_0 \in \Theta$. Let $\mathcal{P}_\Theta = \{\mathrm{P}_\vartheta : \vartheta \in \Theta\}$ and assume that the parametrization is chosen such that

$$u(\mathrm{P}_\vartheta) = \vartheta, \qquad \vartheta \in \Theta,$$

where $u : \mathcal{P}_\Theta \to \Theta$, with $\Theta = \mathcal{Y}$, denotes the functional of interest, as in Equation (9). Consider a scoring rule of the form

$$S^\dagger_\ell(\mathrm{P}_\vartheta, y) = \ell^\dagger(\vartheta, y) = \ell^\dagger(u(\mathrm{P}_\vartheta), y),$$

where $\ell^\dagger$ is not necessarily strictly consistent with respect to u. Consequently, $S^\dagger_\ell : \mathcal{P}_\Theta \times \mathcal{Y} \to \bar{\mathbb{R}}$ need not be strictly locally proper with respect to $(\mathcal{P}_\Theta, u)$.

By tautological equivalence, $\ell^\dagger$ is strictly consistent with respect to the Bayes-act functional it induces, provided the Bayes act exists and is unique. In the latter case, the Bayes-act functional $u^\circ : \mathcal{P}_\Theta \to \Theta$ is defined by

$$u^\circ(\mathrm{P}_{\vartheta_0}) \coloneqq \operatorname*{arg\,min}_{\vartheta \in \Theta} \mathbb{E}_{\mathrm{P}_{\vartheta_0}} \ell^\dagger(\vartheta, Y) = \operatorname*{arg\,min}_{\vartheta \in \Theta} \mathbb{E}_{\mathrm{P}_{\vartheta_0}} S^\dagger_\ell(\mathrm{P}_\vartheta, Y).$$

Equivalently, for any fixed P_{ϑ_0}, the same value minimizes the associated score divergence,

$$u^\circ(\mathrm{P}_{\vartheta_0}) = \arg\min_{\vartheta\in\Theta} \mathbb{D}^\dagger_{S^\dagger_\ell}(\mathrm{P}_{\vartheta_0}\|\mathrm{P}_\vartheta),$$

whenever this divergence is well defined. Consequently, u° is trivially elicitable on $\mathcal{P}_\Theta$, with $\ell^\dagger$ as a strictly consistent scoring function.

For $S^\dagger_\ell$ to be strictly locally proper with respect to $(\mathcal{P}_\Theta, u)$, it is necessary and sufficient that, whenever the true distribution is P_{ϑ_0}, the expected score is uniquely minimized by reporting P_{ϑ_0}. In functional terms, this means that

$$u^\circ(\mathrm{P}_{\vartheta_0}) = u(\mathrm{P}_{\vartheta_0}) \equiv \vartheta_0, \qquad \vartheta_0 \in \Theta. \tag{F.1}$$

Hence, strict local propriety with respect to $(\mathcal{P}_\Theta, u)$ is equivalent to the Bayes-act functional u° and the intended functional u coinciding on $\mathcal{P}_\Theta$.

Indeed, if the two functionals do not coincide, then there exists some $\mathrm{P}_{\vartheta_0} \in \mathcal{P}_\Theta$ such that $u^\circ(\mathrm{P}_{\vartheta_0}) = \vartheta^\circ \neq \vartheta_0$. By definition of the Bayes act, the expected score under P_{ϑ_0} is then uniquely minimized at $\mathrm{P}_{\vartheta^\circ}$ rather than at P_{ϑ_0}. Hence

$$\mathbb{E}_{\mathrm{P}_{\vartheta_0}} S^\dagger_\ell(\mathrm{P}_{\vartheta^\circ}, Y) < \mathbb{E}_{\mathrm{P}_{\vartheta_0}} S^\dagger_\ell(\mathrm{P}_{\vartheta_0}, Y).$$

This violates propriety with respect to the intended functional u, because $u(\mathrm{P}_{\vartheta^\circ}) \neq u(\mathrm{P}_{\vartheta_0})$ while the forecast with the wrong functional value attains a strictly smaller expected score.

Conversely, if $u^\circ(\mathrm{P}_{\vartheta_0}) = u(\mathrm{P}_{\vartheta_0}) = \vartheta_0$ for every $\vartheta_0 \in \Theta$, then, by definition of u°,

$$\mathbb{E}_{\mathrm{P}_{\vartheta_0}} S^\dagger_\ell(\mathrm{P}_{\vartheta_0}, Y) \leq \mathbb{E}_{\mathrm{P}_{\vartheta_0}} S^\dagger_\ell(\mathrm{P}_\vartheta, Y), \qquad \vartheta \in \Theta.$$

Moreover, uniqueness of the Bayes act implies that equality holds if and only if $\vartheta = \vartheta_0$. Since $u(\mathrm{P}_\vartheta) = \vartheta$ on $\mathcal{P}_\Theta$, this is equivalent to

$$u(\mathrm{P}_\vartheta) = u(\mathrm{P}_{\vartheta_0}).$$

Hence $S_\ell^\dagger$ is strictly locally proper with respect to $(\mathcal{P}_\Theta, u)$ by Definition 1.

Reintroducing time, Theorem 1 applies regardless of whether the functional equivalence in Equation (F.1) holds. Consider

$$\phi_{S_\ell^\dagger}(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\psi_{S_\ell^\dagger}(y_t, \vartheta_{t|t-1}).$$

If $\mathbb{E}_{\mathrm{P}_{\vartheta_t}}\psi_{S_\ell^\dagger}(Y_t, \vartheta_{t|t-1}) \neq 0$ and $A_{t|t-1} \succ O_k$, then the update is expected-gradient aligned with $S_\ell^\dagger$. Hence, under the regularity conditions of Theorem 1, there exists $\bar{\kappa} > 0$ such that, for all $\kappa \in (0, \bar{\kappa}]$,

$$\Delta\mathbb{D}^\dagger_{S_\ell^\dagger}\big((\phi_{S_\ell^\dagger})_\kappa\big) < 0.$$

This reduction concerns the score divergence $\mathbb{D}^\dagger_{S_\ell^\dagger}$ and does not require strict local propriety with respect to $(\mathcal{P}_\Theta, u)$. It is therefore directed toward the Bayes-act functional induced by $S_\ell^\dagger$,

$$u^\circ(\mathrm{P}_{\vartheta_t}) = \arg\min_{\vartheta\in\Theta} \mathbb{E}_{\mathrm{P}_{\vartheta_t}} S_\ell^\dagger(\mathrm{P}_\vartheta, Y_t),$$

whenever this Bayes act exists and is unique. Only if Equation (F.1) holds does this target coincide with the intended functional $u(\mathrm{P}_{\vartheta_t})$; only then is the update a PRADA update with respect to $(\mathcal{P}_\Theta, \mathbb{D}_{S_\ell^\dagger}, u)$.

F.2 Misspecification

We generalize the setup and notation of Appendix F.1. Let $\mathcal{P}_\Theta = \{\mathrm{F}_\vartheta : \vartheta \in \Theta\} \subseteq \mathcal{P}$ denote the working class, and $\mathrm{P} \in \mathcal{P}$. The concept of strict propriety as introduced by Gneiting and Raftery (2007) is too narrow to handle cases where $\mathrm{P} \in \mathcal{P}\backslash\mathcal{P}_\Theta$ in the score divergence $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}_\vartheta) \coloneqq \mathbb{E}_\mathrm{P} S(\mathrm{F}_\vartheta, Y) - \mathbb{E}_\mathrm{P} S(\mathrm{P}, Y)$. Indeed, under this restriction, we cannot conclude that S is strictly proper with respect to $\mathcal{P}$ since the expected score comparison evaluates

the score at the true distribution P, even though the admissible forecasts are restricted to $\mathcal{P}_\Theta$.

However, the pseudo-projected scoring rule $S^*_{|\mathcal{Q}} : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$, with working class $\mathcal{Q} \subseteq \mathcal{P}$,

$$S^*(\mathrm{P}, y) \coloneqq S(\Pi_{\mathcal{Q}}^{\mathbb{D}_S}(\mathrm{P}), y),$$

where $\Pi_{\mathcal{Q}}^{\mathbb{D}_S} : \mathcal{P} \to \mathcal{Q}$ denotes the projection of P onto $\mathcal{Q}$ in the geometry of $\mathbb{D}_S$, is strictly locally proper with respect to $(\mathcal{P}, \Pi_{\mathcal{Q}}^{\mathbb{D}_S})$. For $\mathcal{Q} = \mathcal{P}_\Theta$, the projection reduces to the pseudo-true value functional $u^* : \mathcal{P} \to \Theta$,

$$u^*(\mathrm{P}) \coloneqq \arg\min_{\vartheta \in \Theta} \mathbb{E}_{\mathrm{P}} S(\mathrm{F}_\vartheta, Y), \tag{F.2}$$

provided the minimizer exists and is unique. For the given true distribution P, write

$$\vartheta^* \equiv u^*(\mathrm{P}).$$

By construction, the projected divergence identifies ϑ^*, that is,

$$\mathbb{D}_{S^*}\big(\mathrm{P} \| \mathrm{F}_\vartheta\big) = 0 \iff u\big(\Pi_{\mathcal{P}_\Theta}^{\mathbb{D}_S}(\mathrm{F}_\vartheta)\big) = u\big(\Pi_{\mathcal{P}_\Theta}^{\mathbb{D}_S}(\mathrm{P})\big) \iff \vartheta = \vartheta^*.$$

Reintroducing time indices, under the regularity conditions of Corollary 1, the derivative-adaptive updating rule $\phi_S(y_t, \vartheta_{t|t-1}) = \vartheta_{t|t-1} - A_{t|t-1}\psi_S(y_t, \vartheta_{t|t-1})$ is therefore a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_S, u^*)$, regardless of whether $\mathrm{P}_t \in \mathcal{P}_\Theta$. The reduction in score divergence is directed toward $\mathrm{F}_{u^*(\mathrm{P}_t)}$, the projection of P_t onto the working class in the geometry induced by $\mathbb{D}_S$.

The pseudo-true value ϑ^* admits no general agreement with the intended functional value

$$\vartheta_u \coloneqq u(\mathrm{P}) \in \Theta,$$

when $\mathrm{P} \in \mathcal{P} \backslash \mathcal{P}_\Theta$, where $u : \mathcal{P} \to \Theta$, with $\Theta = \mathcal{Y}$, denotes the intended functional, satisfying $u(\mathrm{F}_\vartheta) = \vartheta$ for $\vartheta \in \Theta$. For instance, if u is the mean functional, then ϑ_u is the actual mean

of P, regardless of whether P belongs to the working class. The link between u^* and u is provided directly by Definition 1. Suppose that $S_\ell : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is strictly locally proper with respect to $(\mathcal{P}, u)$. Then, by Definition 1,

$$\mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}_\vartheta, Y) - \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y) \geq 0, \qquad \forall\, \mathrm{F}_\vartheta \in \mathcal{P}_\Theta,$$

with equality if and only if $u(\mathrm{F}_\vartheta) = u(\mathrm{P})$, that is, $\vartheta = u(\mathrm{P})$. Hence $u(\mathrm{P})$ is the unique minimizer of $\vartheta \mapsto \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}_\vartheta, Y)$ over Θ, so that

$$u^*(\mathrm{P}) = u(\mathrm{P}), \qquad \forall\, \mathrm{P} \in \mathcal{P}. \tag{F.3}$$

Equivalently, $\vartheta^* = \vartheta_u$ and the projection preserves the intended functional, $u\big(\Pi^{\mathbb{D}_{S_\ell}}_{\mathcal{P}_\Theta}(\mathrm{P})\big) = u(\mathrm{P})$. Under (F.3), F_{ϑ^*} is the element of $\mathcal{P}_\Theta$ that shares with P the intended functional value, while possibly differing from P in all aspects not captured by u. Misspecification then concerns the distributional shape outside the target functional, not the functional value itself. Consequently, $\mathbb{D}_{S_\ell}$ is a local divergence with respect to u on $\mathcal{P}$, and ϕ_{S_ℓ} is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}, u)$ by Corollary 1.

Now consider a potentially misaligned scoring rule $S^\dagger_\ell : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$, generated by a scoring function through

$$S^\dagger_\ell(\mathrm{F}, y) = \ell^\dagger(u(\mathrm{F}), y), \qquad \mathrm{F} \in \mathcal{P},$$

where $\ell^\dagger$ is not strictly consistent with respect to $(\mathcal{P}, u)$. The Bayes-act functional induced by $S^\dagger_\ell$,

$$u^\circ(\mathrm{P}) \coloneqq \operatorname*{arg\,min}_{\vartheta \in \Theta} \mathbb{E}_{\mathrm{P}} \ell^\dagger(\vartheta, Y).$$

For the given true distribution P, write

$$\vartheta^\circ \equiv u^\circ(\mathrm{P}).$$

In general, $\vartheta^\circ \neq \vartheta_u$, and, under distributional misspecification, also possibly $\vartheta^\circ \neq \vartheta^*$. Hence one has to distinguish between the pseudo-true value

$$\vartheta^* = u^*(\mathrm{P}) = u\big(\Pi^{\mathbb{D}_{S_\ell}}_{\mathcal{P}_\Theta}(\mathrm{P})\big),$$

selected by the intended projected local divergence, the intended functional value

$$\vartheta_u = u(\mathrm{P}),$$

and the score-induced pseudo-target

$$\vartheta^\circ = u^\circ(\mathrm{P}),$$

selected by $S^\dagger_\ell$. The equality $\vartheta^* = \vartheta_u$ holds only when the projection onto the working class preserves the intended functional value, that is, $u\big(\Pi^{\mathbb{D}_{S_\ell}}_{\mathcal{P}_\Theta}(\mathrm{P})\big) = u(\mathrm{P})$.

The scoring rule $S^\dagger_\ell$ is aligned with the intended functional if and only if its Bayes-act functional coincides with u on $\mathcal{P}$, that is, $u^\circ(\mathrm{P}) = u(\mathrm{P}),\ \mathrm{P} \in \mathcal{P}$. In that case, for every $\mathrm{P} \in \mathcal{P}$, the expected score is uniquely minimized by reporting the intended functional value. Hence $S^\dagger_\ell$ is strictly locally proper with respect to $(\mathcal{P}, u)$. Consequently, the score divergence $\mathbb{D}^\dagger_{S^\dagger_\ell}$ is a local divergence with respect to u, which we denote by $\mathbb{D}_{S^\dagger_\ell}$. Under the regularity conditions of Corollary 1, the derivative-adaptive update based on $S^\dagger_\ell$ is therefore a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S^\dagger_\ell}, u)$.

If, instead, $u^\circ(\mathrm{P}) \neq u(\mathrm{P})$ for some $\mathrm{P} \in \mathcal{P}$, then the derivative-adaptive update based on $S^\dagger_\ell$ may still reduce the score divergence induced by $S^\dagger_\ell$. This reduction is directed toward the Bayes act ϑ°, however, and not toward the intended functional value ϑ_u. Hence the update is not a PRADA update with respect to $(\mathcal{P}, \mathbb{D}, u)$ for any local divergence $\mathbb{D}$ identifying u.

G Proofs of the results in Appendix A

G.1 Proof of Proposition 3

Proof. We start with the ' $\Longrightarrow$ ' direction. Assume ℓ is strictly consistent with respect to $(\tilde{u}, \mathcal{P})$. Fix $\mathrm{P}, \mathrm{F} \in \mathcal{P}$. By strict consistency, every $x \in \tilde{u}(\mathrm{P})$ attains the same expected score $\mathbb{E}_{\mathrm{P}}\ell(x, Y) = \ell^*_{\mathrm{P}}$, while $\mathbb{E}_{\mathrm{P}}\ell(x, Y) > \ell^*_{\mathrm{P}}$ for all $x \notin \tilde{u}(\mathrm{P})$. Correspondingly, $\sup_{x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}}) = \ell^*_{\mathrm{P}}$. So, if $\tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F}) \neq \emptyset$, we can always pick $\tilde{x} \in \tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F})$ such that $\mathbb{E}_{\mathrm{P}}\ell(\tilde{x}, Y) = \ell^*_{\mathrm{P}}$. At the same time, any $x_{\mathrm{F}} \notin \tilde{u}(\mathrm{P})$ yields a strictly larger expectation. Hence $\inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}, Y; x_{\mathrm{F}}) = \inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{F}}, Y) = \ell^*_{\mathrm{P}}$ and equality holds. In contrast, if $\tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F}) = \emptyset$, every $x_{\mathrm{F}} \in \tilde{u}(\mathrm{F})$ lies outside $\tilde{u}(\mathrm{P})$ and hence $\inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{F}}, Y) > \ell^*_{\mathrm{P}}$.

For the ' $\Longleftarrow$ ' direction, we assume that S_ℓ is strictly locally proper with respect to $\tilde{u}$. Fix $\mathrm{P}, \mathrm{F} \in \mathcal{P}$. By definition, it holds that

$$\sup_{x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}}) \leq \inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}, Y; x_{\mathrm{F}}).$$

Hence, for any $x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$, $\mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{P}}, Y) \leq \inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{F}}, Y)$. Next, we use the richness assumption pointwise. For arbitrary $x \in \mathcal{Y}$, choose $\mathrm{F}_x \in \mathcal{P}$ such that $\tilde{u}(\mathrm{F}_x) = \{x\}$. Then

$$\mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{P}}, Y) \leq \inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F}_x)} \mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{F}}, Y) = \mathbb{E}_{\mathrm{P}}\ell(x, Y),$$

so every $x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$ is a minimizer of $x \mapsto \mathbb{E}_{\mathrm{P}}\ell(x, Y)$ over $\mathcal{Y}$.

To see that equality occurs only if $x \in \tilde{u}(\mathrm{P})$, consider $\bar{x} \notin \tilde{u}(\mathrm{P})$ and let $\mathrm{F}_{\bar{x}} \in \mathcal{P}$ be the distribution such that $\tilde{u}(\mathrm{F}_{\bar{x}}) = \{\bar{x}\}$. Strict local propriety yields

$$\sup_{x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})} \mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{P}}, Y) < \inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F}_{\bar{x}})} \mathbb{E}_{\mathrm{P}}\ell(x_{\mathrm{F}}, Y) = \mathbb{E}_{\mathrm{P}}\ell(\bar{x}, Y),$$

so $\bar{x}$ cannot attain the minimum. Hence equality holds only for $x \in \tilde{u}(\mathrm{P})$. Consequently, ℓ is strictly consistent with respect to $(\mathcal{P}, \tilde{u})$. This completes the equivalence. □

G.2 Proof of Proposition 4

Proof. We verify the conditions in Definition 2 one by one. Fix a $\mathrm{P}, \mathrm{F}, \mathrm{G} \in \mathcal{P}$.

(i) Assume $\tilde{u}(\mathrm{F}) = \tilde{u}(\mathrm{G})$. Then the infimum is taken over the same set. Moreover, the members of the scoring rule family coincide by condition (i) in Definition 8. Hence,

$$\mathbb{D}_{S_\ell}(\mathrm{P}\|\mathrm{F}; x_{\mathrm{P}}) = \mathbb{D}_{S_\ell}(\mathrm{P}\|\mathrm{G}; x_{\mathrm{P}}), \quad \forall x_{\mathrm{P}} \in \tilde{u}(\mathrm{P}).$$

(ii) By condition (ii) in Definition 8, it follows that for all $x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$, we have that

$$\mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}}) \leq \sup_{x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}}) \leq \inf_{x_{\mathrm{F}} \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}, Y; x_{\mathrm{F}}).$$

Taking the infimum over $x \in \tilde{u}(\mathrm{F})$ of the difference yields

$$\mathbb{D}_{S_\ell}(\mathrm{P}\|\mathrm{F}; x_{\mathrm{P}}) = \inf_{x \in \tilde{u}(\mathrm{F})} \{\mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}, Y; x) - \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}})\} \geq 0.$$

(iii) For the ' $\Longrightarrow$ ' direction, assume $\mathbb{D}_{S_\ell}(\mathrm{P}\|\mathrm{F}; x_{\mathrm{P}}) = 0$ for some $x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$. Then $\inf_{x \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}, Y; x) = \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}})$. Together with (ii), we obtain

$$\mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}}) \leq \sup_{z \in \tilde{u}(\mathrm{P})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; z) \leq \inf_{x \in \tilde{u}(\mathrm{F})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{F}, Y; x) = \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}})$$

so equality holds throughout. By the equality case in condition (ii) of Definition 8 this implies $\tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F}) \neq \emptyset$.

For the ' $\Longleftarrow$ ' direction, suppose $\tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F}) \neq \emptyset$. Applying condition (ii) with $\mathrm{P} = \mathrm{F}$ shows that $\mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}})$ is constant over $x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$. Specifically,

$$\sup_{x \in \tilde{u}(\mathrm{P})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x) = \inf_{x \in \tilde{u}(\mathrm{P})} \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x) = \mathbb{E}_{\mathrm{P}} S_\ell(\mathrm{P}, Y; x_{\mathrm{P}}),$$

for every $x_{\mathrm{P}} \in \tilde{u}(\mathrm{P})$, which yields $\mathbb{D}_{S_\ell}(\mathrm{P}\|\mathrm{F}; x_{\mathrm{P}}) = 0$. Hence $\mathbb{D}_{S_\ell}(\mathrm{P}\|\mathrm{F}; x_{\mathrm{P}}) = 0$ if and only if $\tilde{u}(\mathrm{P}) \cap \tilde{u}(\mathrm{F}) \neq \emptyset$. □

G.3 Proof of Corollary 4

Proof. Fix a $\mathrm{P}_t \in \mathcal{P}$ and $\vartheta_{t|t-1} \in \Theta$. Verify that

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_t}\psi_{S_\ell^{\zeta^*}}(Y_t, \vartheta_{t|t-1})^\top \mathbb{E}_{\mathrm{P}_t}\Delta\phi_{S_\ell^{\zeta^*}}(Y_t, \vartheta_{t|t-1}) \\
&\quad = -\mathbb{E}_{\mathrm{P}_t}\psi_{S_\ell^{\zeta^*}}(Y_t, \vartheta_{t|t-1})^\top A_{t|t-1}\mathbb{E}_{\mathrm{P}_t}\psi_{S_\ell^{\zeta^*}}(Y_t, \vartheta_{t|t-1}) \\
&\quad < 0,
\end{aligned}$$

since $A_{t|t-1} \succ O_k$ and $\mathbb{E}_{\mathrm{P}_t}\psi_{S_\ell^{\zeta^*}}(Y_t, \vartheta_{t|t-1}) \neq 0$ by assumption. Hence, the updating rule $\phi_{S_\ell^{\zeta^*}}$ is EGA for $S_\ell^{\zeta^*}$ at $\vartheta_{t|t-1} \in \Theta$. Applying Theorem 1 with $S = S_\ell^{\zeta^*}$ and $\phi = \phi_{S_\ell^{\zeta^*}}$ yields the existence of a $\bar{\kappa} > 0$ such that for all $0 < \kappa \leq \bar{\kappa}$ it holds that

$$\begin{aligned}
0 &> \Delta\mathbb{D}_{S_\ell^{\zeta^*}}(\phi_{S_\ell^{\zeta^*},\kappa}) \\
&= \mathbb{E}_{\mathrm{P}_t}\mathbb{D}_{S_\ell^{\zeta^*}}(\mathrm{P}_t \| \mathrm{F}^\kappa_{t|t}) - \mathbb{E}_{\mathrm{P}_t}\mathbb{D}_{S_\ell^{\zeta^*}}(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) \\
&\geq \int_{\mathcal{Y}} \inf_{x \in \tilde{u}(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)})} \mathbb{E}_{\mathrm{P}_t} S_\ell(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)}, Y_t; x) - \mathbb{E}_{\mathrm{P}_t} S_\ell^{\zeta^*}(\mathrm{F}_{t|t-1}, Y_t)\mathrm{P}_t(\mathrm{d}y).
\end{aligned}$$

For the argmin selector ζ^*, is attained pointwise, so

$$\begin{aligned}
&\int_{\mathcal{Y}} \inf_{x \in \tilde{u}(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)})} \mathbb{E}_{\mathrm{P}_t} S_\ell(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)}, Y_t; x)\mathrm{P}_t(\mathrm{d}y) \\
&\qquad = \int_{\mathcal{Y}} \mathbb{E}_{\mathrm{P}_t} S_\ell\left(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)}, Y_t; \zeta^* \circ \tilde{u}(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)})\right) \mathrm{P}_t(\mathrm{d}y),
\end{aligned}$$

and, moreover,

$$\begin{aligned}
0 &> \int_{\mathcal{Y}} \mathbb{E}_{\mathrm{P}_t} S_\ell\left(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)}, Y_t; \zeta^* \circ \tilde{u}(\mathrm{F}_{\vartheta^\kappa_{t|t}(y)})\right) - \mathbb{E}_{\mathrm{P}_t} S_\ell\left(\mathrm{F}_{t|t-1}, Y_t; \zeta^* \circ \tilde{u}(\mathrm{F}_{t|t-1})\right) \mathrm{P}_t(\mathrm{d}y) \\
&= \mathbb{E}_{\mathrm{P}_t}\mathbb{D}_{S_\ell}(\mathrm{P}_t \| \mathrm{F}_{t|t}; x_{\mathrm{P}_t}) - \mathbb{E}_{\mathrm{P}_t}\mathbb{D}_{S_\ell}(\mathrm{P}_t \| \mathrm{F}_{t|t-1}; x_{\mathrm{P}_t}),
\end{aligned}$$

where $\mathbb{D}_{S_\ell}(\mathrm{P}_t \| \mathrm{F}_{t|t}; x_{\mathrm{P}})$ is a local divergence by Proposition 4, for all $x_{\mathrm{P}_t} \in \tilde{u}(\mathrm{P}_t)$. Hence, $\phi_{S_\ell^{\zeta^*},\kappa}$ is $\mathbb{D}_{S_\ell}(\cdot\|\cdot; x_{\mathrm{P}_t})$-reducing for all $x_{\mathrm{P}_t} \in \tilde{u}(\mathrm{P}_t)$, $\kappa \in (0, \bar{\kappa}]$. Therefore, $\phi_{S_\ell^{\zeta^*}}(y_t, \vartheta_{t|t-1})$ is a PRADA update with respect to $(\mathcal{P}, \mathbb{D}_{S_\ell}(\cdot\|\cdot; \zeta^* \circ \tilde{u}(\cdot)), \tilde{u})$. □

H Derivations Appendix E

H.1 Example E.2.1

We first define $\bar{z}_j := \frac{r_j - \mu_{t|t-1}}{\sigma_{t|t-1}}$ for $j \in \{1, 2\}$. For the censoring mass $\bar{F}_{t|t-1}$, we obtain $\bar{F}_{t|t-1} = \Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)$. For the mean part of the score, we note that

$$\begin{aligned}
&\left[\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} s(Y_t, \vartheta_{t|t-1})\right]_1 \\
&\quad = -\frac{1}{\bar{F}_{t|t-1}} \left(\int_{-\infty}^{r_1} \frac{y - \mu_{t|t-1}}{\sigma^2_{t|t-1}} f_{t|t-1}(y) \mathrm{d}y + \int_{r_2}^{\infty} \frac{y - \mu_{t|t-1}}{\sigma^2_{t|t-1}} f_{t|t-1}(y) \mathrm{d}y \right) \\
&\quad = -\frac{1}{\bar{F}_{t|t-1}\sigma_{t|t-1}} \left(\int_{-\infty}^{\bar{z}_1} z\varphi(z) \mathrm{d}z + \int_{\bar{z}_2}^{\infty} z\varphi(z) \mathrm{d}z \right) \\
&\quad = \frac{1}{\sigma_{t|t-1}} \frac{\varphi(\bar{z}_1) - \varphi(\bar{z}_2)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)},
\end{aligned}$$

where $[a]_j$ denotes the j-th element of the vector a. For the variance part, we first compute

$$\begin{aligned}
\int_{-\infty}^{\bar{z}_1} (1 - z^2)\varphi(z) \mathrm{d}z &= \Phi(\bar{z}_1) - \int_{-\infty}^{\bar{z}_1} z^2 \varphi(z) \mathrm{d}z \\
&= \Phi(\bar{z}_1) - (\Phi(\bar{z}_1) - \bar{z}_1\varphi(\bar{z}_1)) \\
&= \bar{z}_1\varphi(\bar{z}_1),
\end{aligned}$$

and similarly $\int_{\bar{z}_2}^{\infty} (1 - z^2)\varphi(z) \mathrm{d}z = -\bar{z}_2\varphi(\bar{z}_2)$. Therefore,

$$\left[\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} s(Y_t, \vartheta_{t|t-1})\right]_2 = \frac{1}{2} \frac{\bar{z}_1\varphi(\bar{z}_1) - \bar{z}_2\varphi(\bar{z}_2)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)}.$$

Collecting terms, we find

$$\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} s(Y_t, \vartheta_{t|t-1}) = \begin{pmatrix} \frac{1}{\sigma_{t|t-1}} \frac{\varphi(\bar{z}_1) - \varphi(\bar{z}_2)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} \\ \frac{1}{2} \frac{\bar{z}_1\varphi(\bar{z}_1) - \bar{z}_2\varphi(\bar{z}_2)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} \end{pmatrix}.$$

For the Hessian

$$H(y_t, \vartheta_{t|t-1}) = \begin{pmatrix} \frac{1}{\sigma^2_{t|t-1}} & \frac{z_t}{\sigma_{t|t-1}} \\ \frac{z_t}{\sigma_{t|t-1}} & \frac{1}{2} z_t^2 \end{pmatrix},$$

we first note that $\left[\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} H(Y_t, \vartheta_{t|t-1})\right]_{11} = \frac{1}{\sigma^2_{t|t-1}}$.

For the off-diagonals, we obtain

$$\begin{aligned}\left[\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} H(Y_t, \vartheta_{t|t-1})\right]_{12} &= \frac{1}{\bar{F}_{t|t-1}} \left(\int_{-\infty}^{\bar{z}_1} \frac{z}{\sigma_{t|t-1}} \varphi(z) \mathrm{d}z + \int_{\bar{z}_2}^{\infty} \frac{z}{\sigma_{t|t-1}} \varphi(z) \mathrm{d}z \right) \\ &= \frac{1}{\sigma_{t|t-1}} \frac{\varphi(\bar{z}_2) - \varphi(\bar{z}_1)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)}.\end{aligned}$$

And finally,

$$\begin{aligned}&\left[\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} H(Y_t, \vartheta_{t|t-1})\right]_{22} \\ &\quad = \frac{1}{\bar{F}_{t|t-1}} \left(\int_{-\infty}^{\bar{z}_1} \frac{1}{2} z^2 \varphi(z) \mathrm{d}z + \int_{\bar{z}_2}^{\infty} \frac{1}{2} z^2 \varphi(z) \mathrm{d}z \right) \\ &\quad = \frac{1}{2\bar{F}_{t|t-1}} \left(\left(\Phi(\bar{z}_1) - \bar{z}_1 \varphi(\bar{z}_1) \right) + \left(1 - \Phi(\bar{z}_2) + \bar{z}_2 \varphi(\bar{z}_2) \right) \right) \\ &\quad = \frac{1}{2} \left(1 + \frac{\bar{z}_2 \varphi(\bar{z}_2) - \bar{z}_1 \varphi(\bar{z}_1)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} \right).\end{aligned}$$

Collecting entries,

$$\mathbb{E}_{\mathrm{F}^\sharp_{1-w}} H(Y_t, \vartheta_{t|t-1}) = \begin{pmatrix} \frac{1}{\sigma^2_{t|t-1}} & \frac{1}{\sigma_{t|t-1}} \frac{\varphi(\bar{z}_2) - \varphi(\bar{z}_1)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} \\ \frac{1}{\sigma_{t|t-1}} \frac{\varphi(\bar{z}_2) - \varphi(\bar{z}_1)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} & \frac{1}{2} \left(1 + \frac{\bar{z}_2 \varphi(\bar{z}_2) - \bar{z}_1 \varphi(\bar{z}_1)}{\Phi(\bar{z}_1) + 1 - \Phi(\bar{z}_2)} \right) \end{pmatrix}.$$

References

Allen, S., D. Ginsbourger, and J. Ziegel (2023), "Evaluating forecasts for high-impact events using transformed kernel scores", *SIAM/ASA Journal on Uncertainty Quantification*, *11*(3), 906–940.

Amisano, G. and R. Giacomini (2007), "Comparing density forecasts via weighted likelihood ratio tests", *Journal of Business & Economic Statistics*, *25*(2), 177–190.

Barron, J. T. (2019). "A general and adaptive robust loss function". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 4326–4334.

Basel Committee on Banking Supervision (2010). "Basel III: A global regulatory framework for more resilient banks and banking systems". Technical report, Bank for International Settlements, Basel, Switzerland.

Blasques, F., C. Francq, and S. Laurent (2023), "Quasi score-driven models", *Journal of Econometrics*, *234*(1), 251–275.

Blasques, F., S. J. Koopman, and A. Lucas (2015), "Information-theoretic optimality of observation-driven time series models for continuous responses", *Biometrika*, *102*(2), 325–343.

Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasticity", *Journal of Econometrics*, *31*(3), 307–327.

Catania, L., A. C. Harvey, and A. Luati (2025), "Robust CDF-Filtering of a Location Parameter", *Journal of Time Series Analysis*, , jtsa.70026.

Catania, L. and A. Luati (2023), "Semiparametric modeling of multiple quantiles", *Journal of Econometrics*, *237*(2), 105365.

Cox, D. R. (1981), "Statistical analysis of time series: Some recent developments", *Scandinavian Journal of Statistics*, *8*(2), 93–115.

Creal, D., S. J. Koopman, and A. Lucas (2013), "Generalized autorgressive score models with applications", *Journal of Applied Econometrics*, *28*(5), 777–795.

Creal, D., S. J. Koopman, A. Lucas, and M. Zamojski (2024), "Observation-driven filtering of time-varying parameters using moment conditions", *Journal of Econometrics*, *238*(2), 105635.

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

Dawid, A. P. and M. Musio (2014), "Theory and applications of proper scoring rules", *METRON*, *72*(2), 169–183.

De Punder, R., T. Dimitriadis, and R.-J. Lange (2026). "Kullback-Leibler-based characterizations of score-driven updates". Version Number: 2.

De Punder, R. F. A., C. G. H. Diks, R. J. A. Laeven, and D. J. C. Van Dijk (2025), "Localizing strictly proper scoring rules", *Journal of the American Statistical Association*, , 1–24.

Diks, C., V. Panchenko, and D. Van Dijk (2011), "Likelihood-based scoring rules for comparing density forecasts in tails", *Journal of Econometrics*, *163*(2), 215–230.

Dimitriadis, T., T. Fissler, and J. Ziegel (2024), "Osband's principle for identification functions", *Statistical Papers*, *65*(2), 1125–1132.

Donker van Heel, S., R.-J. Lange, D. J. Van Dijk, and B. Van Os (2025). "Stability and performance guarantees for misspecified multivariate score-driven filters". Tinbergen Instiute Discussion Paper 25-006/III.

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

Engle, R. F. (1982), "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation", *Econometrica*, *50*(4), 987–1007.

Fissler, T. and J. F. Ziegel (2016), "Higher order elicitability and Osband's principle", *The Annals of Statistics*, *44*(4), 1680–1707.

Gneiting, T. (2011), "Making and evaluating point forecasts", *Journal of the American Statistical Association*, *106*(494), 746–762.

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

Gneiting, T. and R. Ranjan (2011), "Comparing density forecasts using threshold- and quantile-weighted scoring rules", *Journal of Business & Economic Statistics*, *29*(3), 411–422.

Good, I. J. (1952), "Rational decisions", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *14*(1), 107–114.

Gorgi, P. (2020), "Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(5), 1325–1347.

Gorgi, P., C. S. A. Lauria, and A. Luati (2024), "On the optimality of score-driven models", *Biometrika*, *111*(3), 865–880.

Grünwald, P. D. and A. P. Dawid (2004), "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory", *The Annals of Statistics*, *32*(4).

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators", *Econometrica*, *50*(4), 1029–1054.

Harvey, A., S. Hurn, D. Palumbo, and S. Thiele (2024), "Modelling circular time series", *Journal of Econometrics*, *239*(1), 105450.

Harvey, A. and R. Ito (2020), "Modeling time series when some observations are zero", *Journal of Econometrics*, *214*(1), 33–45.

Harvey, A. and Y. Liao (2023), "Dynamic Tobit models", *Econometrics and Statistics*, *26*, 72–83.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Number 52 in Econometric Society Monographs. Cambridge University Press.

Heinrich, C. (2014), "The mode functional is not elicitable", *Biometrika*, *101*(1), 245–251.

Heinrich-Mertsching, C. and T. Fissler (2022), "Is the mode elicitable relative to unimodal distributions?", *Biometrika*, *109*(4), 1157–1164.

Holzmann, H. and B. Klar (2017), "Focusing on regions of interest in forecast evaluation", *The Annals of Applied Statistics*, *11*(4).

Huber, P. J. (1964), "Robust estimation of a location parameter", *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

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

Kolda, T. G. and B. W. Bader (2009), "Tensor decompositions and applications", *SIAM Review*, *51*(3), 455–500.

Kullback, S. and R. A. Leibler (1951), "On information and sufficiency", *The Annals of Mathematical Statistics*, *22*(1), 79–86.

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

Neyman, J. and E. Pearson (1933), "IX. On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337.

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

Patton, A. J., J. F. Ziegel, and R. Chen (2019), "Dynamic semiparametric models for expected shortfall (and Value-at-Risk)", *Journal of Econometrics*, *211*(2), 388–413.

Pelenis, J. (2014). "Weighted scoring rules for comparison of density forecasts on subsets of interest". Available at `https://web.archive.org/web/20170909102128/https://elaine.ihs.ac.at/~pelenis/JPelenis_wsr.pdf`.

Selten, R. (1998), "Axiomatic characterization of the quadratic scoring rule", *Experimental Economics*, *1*(1), 43–62.

Steinwart, I. and J. F. Ziegel (2021), "Strictly proper kernel scores and characteristic kernels on compact spaces", *Applied and Computational Harmonic Analysis*, *51*, 510–542.

Van den Boomgaard, R. and J. Van de Weijer (2002). "On the equivalence of local-mode finding, robust estimation and mean-shift analysis as used in early vision tasks". In *2002 International Conference on Pattern Recognition*, Volume 3, pp. 927–930.

Van der Vaart, A. W. (1998). *Asymptotic Statistics* (1st ed.). Cambridge University Press.

Wald, A. (1949), "Note on the consistency of the maximum likelihood estimate", *The Annals of Mathematical Statistics*, *20*(4), 595–601.