TI 2026-023/III
Tinbergen Institute Discussion Paper

Barron-Loss Adaptive Estimation

Revision: June 2026

Ramon F. A. de Punder[1]
Mathijs R. G. Dijkstra[2]
Cees G. H. Diks[3]

1 University of Amsterdam and Tinbergen Institute

2 University of Amsterdam and Tinbergen Institute

3 University of Amsterdam and Tinbergen Institute

Barron-Loss Adaptive Estimation*

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

Mathijs R. G. Dijkstra
Department of Quantitative Economics
University of Amsterdam and Tinbergen Institute

Cees G. H. Diks
Department of Quantitative Economics
University of Amsterdam and Tinbergen Institute

June 5, 2026

Abstract

The score-driven framework relies on pre-specified scoring rules tied to assumed conditional densities, making it vulnerable to misspecification under outliers or structural breaks. We embed the flexible Barron loss within the quasi score-driven (QSD) framework, allowing the degree of robustness to be learned from the data. The resulting Barron-Loss Adaptive Estimation (BLADE) filter generates a strictly stationary, ergodic, and invertible sequence of time-varying parameters under mild regularity conditions. Within an extended quasi score-driven estimation framework, obtained by generalizing the required moment condition, the associated estimator is shown to be consistent and asymptotically normal. The Barron loss is strictly consistent for a family of functionals indexed by the shape parameter γ, enabling smooth adaptation between classical and robust targets. We establish that the BLADE update belongs to the class of Proper and Robust Autoregressive Derivative Adaptive (PRADA) models and is therefore expected divergence reducing. We further establish that more robust updates achieve an at least as large expected local divergence reduction as less robust ones over explicit intervals, of step sizes in a non-contaminated setting and of contamination proportions under a contaminated updating draw, providing a formal robustness guarantee against corrupted observations. Monte Carlo experiments under non-contaminated and contaminated estimation environments confirm these theoretical findings and demonstrate superior performance relative to GARCH and β_t–GARCH models when outliers are present.

Keywords: Robust statistics, Score-driven models, Local Divergences, Consistent scoring functions, Online Z-Estimation

*We are very grateful to Timo Dimitriadis, Andrew Harvey, Floris Holstege, Rutger-Jan Lange, Sébastien Laurent, Dario Palumbo, Andrew Patton, Dick van Dijk and participants of the Score-Driven and Nonlinear Time Series Models conference (Venice, 2026) and the Nordic Econometric Meeting 2026 for their insightful 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

Over the last decade, score-driven models have emerged as a prominent subclass of observation driven time series models. The use of the gradient of the model-implied log-likelihood, that is, the *score*, to update time-varying parameters was originally proposed by Creal et al. (2013) and Harvey (2013). This framework has since generated a broad range of applications and has been adopted in more than 400 published papers. An up-to-date overview is available at `https://www.gasmodel.com`. For comprehensive surveys, see Artemova et al. (2022a), Artemova et al. (2022b) and Harvey (2022). Interest in the theoretical properties of score-driven models has increased more recently, with contributions by Gorgi et al. (2024), De Punder et al. (2026), and Donker van Heel et al. (2025). In parallel, several generalizations of the original score-driven formulations have been explored (Patton et al. 2019; Catania and Luati 2023; Blasques et al. 2023; Catania et al. 2025; De Punder 2026; Creal et al. 2024).

In this paper, we address concerns raised by Donker van Heel et al. (2025) and Lange et al. (2026) that score-driven filters may become unstable, even in settings as common as volatility filtering under Gaussian assumptions. As suggested by Lange et al. (2026), regularization of the update may improve robustness, in their case arising from an implicit updating mechanism. We take a different approach by retaining the explicit and computationally convenient update of the original score-driven framework while modifying the gradient of the log-likelihood. Specifically, we replace the log-likelihood with the generalized and smoothed Huber loss proposed by Barron (2019) (extended in Barron (2025)), thereby anchoring the update in the robust statistics literature (Huber 1981; Hampel et al. 1986). Related approaches to robust filtering under contamination include Muler and Yohai

(2008), Muler et al. (2009), and Muler and Yohai (2013), who Huberize the innovation inside the residual recursion through a bounded function and combine this with a robust loss in the estimation step, and Calvet et al. (2015), who impose a pointwise worst-case bound on the filtering error under additive perturbations. In all cases the degree of Huberization is fixed ex ante, whereas here it is governed by a single shape parameter learned from the data.

We show that robust updating can reduce the expected loss under less robust evaluation criteria in the presence of outliers or when learning rates are relatively large. For example, median-based updates may improve, relative to squared loss updates, the expected squared loss performance, even though squared loss is strictly consistent for the mean rather than the median (Gneiting 2011).

The recursive Barron-Loss Adaptive Estimation (BLADE) of the time-varying parameter of interest constitutes a special case of both the quasi score-driven (QSD) framework of Blasques et al. (2023) and the Proper and Robust Autoregressive Derivative Adaptive (PRADA) models proposed by De Punder (2026). The former embedding is exploited to establish stationarity and ergodicity, as well as uniform invertibility, which in turn facilitates consistent estimation of static parameters. In doing so, we extend the QSD estimation framework (Blasques et al. 2023, Theorem 1) by generalizing its required moment condition. The nesting within the PRADA class shows that BLADE updates reduce an expected local divergence, closely analogous to score-driven models, where updates reduce the global expected Kullback–Leibler (KL) divergence from the true distribution. In particular, we show that robust updates lead to enhanced reductions toward the true distribution under a given expected local divergence measure. In related work, Creal et al. (2024) propose updates based on the gradient of an expected generalized method of moments criterion (EGMM) and show that these updates improve the EGMM criterion. De Punder et al. (2026), however, emphasize that the conditions under which such improvements hold are

unverifiable in practice. We therefore adopt the PRADA perspective, which links updates directly to reductions in an expected local divergence measure.

In a recent paper, Catania et al. (2025) replace the gradient of the log-likelihood, or, equivalently, the logarithmic scoring rule of Good (1952) by the Continuously Ranked Probability Score (CRPS) of Matheson and Winkler (1976), thereby proposing a robust alternative to classical score-driven updates. Both scoring rules are strictly proper (Gneiting and Raftery 2007), implying that the expected scoring rule difference defines a divergence measure, which reduces to the Kullback–Leibler divergence for the logarithmic scoring rule and to the Cramér distance for the CRPS. Updates based on strictly proper scoring rules are examples of PRADA updates, guaranteed to reduce the respective divergences from the true distribution in expectation (De Punder 2026). The key distinction between the two rules lies in their robustness properties. While the logarithmic scoring rule is well-known for its sensitivity to outliers (Selten 1998), the CRPS has gained prominence due to its inherent robustness. From an estimation perspective, the choice between the logarithmic scoring rule and the CRPS reflects a trade-off between efficiency and robustness. As noted by Gneiting and Raftery (2007, Section 9), when estimating the location parameter of a Gaussian family with known variance, the CRPS reduces to a smooth analogue of the Huber loss, further highlighting its close connection to robust statistics. De Punder (2026, Example 5) likewise considers the CRPS as a robust alternative in this setting and shows that, in contrast to score-driven updates, CRPS-based updates yield uniformly bounded gradients and Hessians, together with the same Winsorizing behavior as the Huber statistic.

Winsorization may alternatively be implemented by applying it directly to the logarithmic scoring rule (Harvey and Ito 2020; Harvey and Liao 2023), a form of Huberizing the update already noted above. Winsorization is closely connected to censoring (Bernoulli 1760; Tobin 1958) and to the corresponding censored likelihood score introduced by Diks et al. (2011). As emphasized by Harvey and Ito (2020), for fat-tailed distributions such

as the generalized beta distribution of the second kind, the resulting winsorized score for location and scale naturally downweights extreme observations, thereby reducing the influence of outliers on the evolution of the filter. Winsorized scoring rules, including those beyond the logarithmic scoring rule, are strictly locally proper with respect to the Winsorizing functional. Moreover, their expected scoring rule differences correspond to local divergences with respect to the same functional (De Punder et al. 2025; De Punder 2026).

Patton et al. (2019) propose a score-driven updating rule for Value-at-Risk and Expected Shortfall, which form a jointly elicitable functional (Fissler and Ziegel 2016). A strictly consistent scoring function exists for this functional, and this is equivalent to the associated induced scoring rule being strictly locally proper with respect to the same functional (Gneiting 2011; De Punder 2026). This unifying perspective is helpful because it directly establishes that updates based on the induced strictly locally proper scoring rule are PRADA. While Patton et al. (2019) motivate their updating rule intuitively as online M-estimation, indications of alternative updating schemes based on Z-estimation are hinted at by Dimitriadis et al. (2024). In a univariate context, the link between online M- and Z-estimation is trivial by Osband‘s principle (Fissler and Ziegel 2016; De Punder 2026).

In contrast to the emphasis on tail-related functionals in Patton et al. (2019), the focus here is on elicitable location functionals of the distribution, such as a conditional mean and median. To remain within the scope of Barron (2019), attention is restricted to univariate time series. The Barron loss encompasses a broad class of loss functions, including the Cauchy or Lorentzian, Geman–McClure, Welsch or Leclerc, generalized Charbonnier, Charbonnier or pseudo-Huber or L^1–L^2, and L^2 loss functions. For the L^2 loss, PRADA updates reduce to well-known univariate filters, such as Autoregressive Moving Average (ARMA) models for the conditional mean and Autoregressive Conditional Heteroscedasticity and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models for

the conditional variance (Engle 1982; Bollerslev 1986). As opposed to score-driven filters, ARMA- and GARCH-type filters are moment-driven rather than likelihood-driven, and the functional form of the filter is correspondingly not implied by the model density. Under correct model specification, this may be associated with a loss of efficiency, since not all available information is incorporated into the filter. In practice, however, model specifications are often incorrect, and filters based on moment conditions rather than the full likelihood may be less sensitive to misspecification. An alternative perspective is that moment driven filters effectively localize the full distribution to a single real-valued functional. A direct consequence of this reduction is that a full likelihood specification becomes redundant. The remaining link to the model likelihood in our setup naturally facilitates improvements toward the true distribution, viewed through the lens of the functional of interest.

The Barron loss is smooth in its robustness parameter and, as illustrated above, highly flexible in terms of the attainable level of robustness. The robustness parameter is to be treated as a tuning parameter optimized using an out-of-sample rolling window loss; this means the required degree of robustness is data-driven. This flexibility is particularly appealing in an online estimation setting. We show that, even when interest is restricted to mean squared error performance within an ARMA-GARCH framework, updating schemes that are more robust than those implied by squared error loss may yield further reductions in expected squared error under data generating processes contaminated with outliers. Accordingly, the Barron loss implied BLADE filter provides a robust alternative to traditional ARMA-GARCH filters when demanded by the data, while retaining these classical filters as special cases under the squared error nesting of the Barron loss.

The remainder of this paper is set up as follows. In Section 2 the formulation of the BLADE filter is proposed along with some theoretical properties of the loss function. Section 3 gives theoretical details on the robustness of the BLADE filter in both a non-

contaminated updating setting and a contaminated updating setting. Section 4 contains details on the asymptotic consistency and normality of the parameters in the BLADE filter. Section 5 outlines the setup of the Monte Carlo study and presents the corresponding results, supplemented by graphical illustrations. Proofs and additional material are collected in the Appendix. The replication code is available at `https://github.com/Mrdijkst/BLADE-Filter`.

2 Barron–loss–based filtering

Consider a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable space $(\mathcal{Y}, \mathcal{G})$, with observation space $\mathcal{Y}$ and $\mathcal{G} \coloneqq \mathcal{B}(\mathcal{Y})$, the Borel σ-algebra on $\mathcal{Y}$. Let $\{Y_t\}_{t\in\mathbb{Z}}$ denote a real-valued stochastic process defined on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathcal{Y} \subseteq \mathbb{R}$. The process is adapted to its natural filtration $\{\mathcal{F}_t^Y\}_{t\in\mathbb{Z}}$, with $\mathcal{F}_t^Y = \sigma(Y_s : s \leq t)$. We condition the true distribution and model distribution on a potentially larger information set $\mathcal{F}_{t-1} \supseteq \mathcal{F}_{t-1}^Y$, assuming Y_t is not $\mathcal{F}_{t-1}$-measurable. Specifically, for each $t \in \mathbb{Z}$, let $\mathrm{P}_t(\cdot)$ denote the true conditional distribution of Y_t given $\mathcal{F}_{t-1}$, defined as the probability kernel $\mathrm{P}_t(\cdot) \coloneqq \mathbb{P}(Y_t \in \cdot \mid \mathcal{F}_{t-1})$, on $(\mathcal{Y}, \mathcal{G})$, with distribution function P_t, and whenever a dominating measure μ_t exists, also μ_t-density $p_t(y) \coloneqq \frac{\mathrm{dP}_t}{\mathrm{d}\mu_t}(y)$.

Model distributions are similarly denoted $\mathrm{F}_{t|t-1}(\cdot) \equiv \mathrm{F}(\cdot \mid \vartheta_{t|t-1})$, with distribution functions $F_{t|t-1}(\cdot)$ and μ_t-densities $f_{t|t-1}(y) \coloneqq \frac{\mathrm{dF}_{t|t-1}}{\mathrm{d}\mu_t}(y)$. The model distribution is conditional on a time-varying parameter estimate $\vartheta_{t|t-1}$, which is assumed to be $\mathcal{F}_{t-1}$-measurable. After observing y_t, the time-varying parameter (and hence distribution) is updated via the *update rule*, $\phi : \mathcal{Y} \times \Theta \to \Theta$, where $\Theta \subseteq \mathbb{R}$ is the open and convex parameter space. This update determines how the predicted parameters $\vartheta_{t|t-1}$ are updated to $\vartheta_{t|t} = \phi(y_t, \vartheta_{t|t-1})$. The update depends on static parameters $\lambda_1 \in \Lambda_1 \subseteq \mathbb{R}^{q_1}$, while notationally suppressed. The update step is followed by a *prediction step*, which yields $\vartheta_{t+1|t}$, constructed as $\vartheta_{t+1|t} = \tilde{\omega} + \tilde{\beta}\vartheta_{t|t}$

for some real-valued parameters $\tilde{\omega}$ and $\tilde{\beta}$, collected in $\lambda_2 \in \Lambda_2 \subseteq \mathbb{R}^{q_2}$, with $q_2 \in \mathbb{N}$ allowing for higher order lags. The model distribution is also allowed to depend on static nuisance parameters, which are collected in $\lambda_3 \in \Lambda_3 \subseteq \mathbb{R}^{q_3}$. The estimation problem is considered for the whole vector $\lambda \coloneqq (\lambda_1^\top, \lambda_2^\top, \lambda_3^\top)^\top \in \Lambda \in \mathbb{R}^q$, with $q = q_1 + q_2 + q_3$.

2.1 PRADA updates for functionals

The BLADE filter of Section 2.2 is a special case of the Proper and Robust Autoregressive Derivative Adaptive (PRADA) framework of De Punder (2026), which extends score-driven updates (Creal et al. 2013; Harvey 2013) by replacing the logarithmic scoring rule with any *strictly locally proper* scoring rule S, thereby generalizing the KL-divergence contraction to the *score divergence* induced by S.

The contraction guarantee is formalized through a score divergence measuring the expected improvement in fit when the model distribution is updated. Let $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$, be a scoring rule, where $\mathcal{P}$ denotes a class of probability measures on $(\mathcal{Y}, \mathcal{G})$. For any such S, the associated *score divergence* is

$$\mathbb{D}_S^\dagger : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \cup \{\infty\}, \qquad \mathbb{D}_S^\dagger(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) \coloneqq \mathbb{E}_{\mathrm{P}_t}\big[S(\mathrm{P}_t, Y_t) - S(\mathrm{F}_{t|t-1}, Y_t)\big]. \tag{1}$$

Following De Punder (2026, Definition 1), S is *strictly locally proper* with respect to $(\mathcal{P}, \mathcal{T})$, for a functional $\mathcal{T}$ on $\mathcal{P}$ if the score divergence is non-negative and zero if and only if $\mathcal{T}(\mathrm{P}_t) = \mathcal{T}(\mathrm{F}_{t|t-1})$ for all $\mathrm{P}_t, \mathrm{F}_{t|t-1} \in \mathcal{P}$. If $\mathbb{D}_S^\dagger$ is based on a strictly locally proper scoring rule it constitutes a *local divergence* $\mathbb{D}_S$ (De Punder 2026, Definition 2) and the triple $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$ a *local divergence space*. When $\mathcal{T} = \mathrm{id}$, where id is the identity functional, strict local propriety reduces to standard strict propriety (Gneiting and Raftery 2007).

This paper specializes to $\mathcal{T} = u$, where $u : \mathcal{P} \to \Theta$ is a robustified k-th moment functional indexed by a tuning parameter, and S is strictly locally proper with respect to

$(\mathcal{P}, u)$. The associated location estimators can, under mild regularity, be formulated as zero-estimators (Z-estimators) satisfying moment conditions of the form

$$\mathbb{E}_{\mathrm{P}_t}\left[\tilde{\psi}\Big(Y_t^k - \vartheta\Big)\right] = 0, \tag{2}$$

where Y_t^k denotes the k-th power of the observation (e.g., $k = 1$ or $k = 2$) and $\tilde{\psi}$ is monotone and odd around zero (Van der Vaart 1998, pp. 42–43). By Osband's principle (Dimitriadis et al. 2024), the identification function $\tilde{\psi}$ in (2) coincides, up to a positive P_t-measurable scalar, with the gradient of the strictly locally proper scoring function S for $(\mathcal{P}, u)$; consequently $\vartheta = u(\mathrm{P}_t)$ is the unique solution to (2) whenever u is single-valued at P_t. As shown in De Punder (2026, Corollary 3), embedding such an identification function in a filter recursion yields a PRADA update with respect to the local divergence space $(\mathcal{P}, \mathbb{D}_S, u)$.

The localization to u is encoded in

$$\vartheta_{t|t-1} \coloneqq u(\mathrm{F}_{t|t-1}), \tag{3}$$

so recursive prediction depends on $\mathrm{F}_{t|t-1}$ only through its local functional value $u(\mathrm{F}_{t|t-1})$. Since $\mathbb{D}_S(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) = 0$ if and only if $u(\mathrm{P}_t) = u(\mathrm{F}_{t|t-1})$, the full specification of $\mathrm{F}_{t|t-1}$ is redundant beyond (3) (Gneiting and Raftery 2007; Gneiting 2011; De Punder 2026). This contrasts with ordinary score-driven models, which account for every feature of the distribution and are therefore sensitive to distributional misspecification. Remaining within the PRADA class guarantees reduction of $\mathbb{D}_S$ for sufficiently small learning rates, regardless of misspecification outside u.

Formally, consider an update rule $\phi : \mathcal{Y} \times \Theta \to \Theta$, indexed by a learning rate or step-size parameter $\tilde{\alpha} \in \mathbb{R}$, which maps the past parameter forecast and the new observation to an updated parameter value, $\vartheta_{t|t} = \phi(y_t, \vartheta_{t|t-1})$. The reduction in the expected divergence

of the update rule ϕ, when updated under a probability measure R_t defined on $(\mathcal{Y}, \mathcal{G})$, is quantified by the *expected divergence reduction*. This reduction serves as a primary criterion for determining update quality, generalizing De Punder (2026, Definition 3) to updates evaluated under R_t that do not necessarily coincide with the true data-generating process P_t.

Definition 1. *Let $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ be a strictly locally proper scoring rule with induced local divergence $\mathbb{D}_S : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$. Let $Y_t \sim \mathrm{P}_t$ and let $\check{Y}_t \sim \mathrm{R}_t$ be the draw used in the update rule $\phi : \mathcal{Y} \times \Theta \to \Theta$. The* expected divergence reduction *is given by*

$$\Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi) \coloneqq \mathbb{D}_S(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) - \mathbb{E}_{\check{Y}_t \sim \mathrm{R}_t}\big[\mathbb{D}_S(\mathrm{P}_t \| \mathrm{F}_{t|t})\big],$$

where the updated model $\mathrm{F}_{t|t} \equiv \mathrm{F}_{\vartheta_{t|t}}$ depends on $\check{Y}_t$ through $\vartheta_{t|t} = \phi(\check{Y}_t, \vartheta_{t|t-1})$.

A positive value of $\Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi)$ indicates that, on average under R_t, the update brings the model distribution closer to P_t in the local divergence $\mathbb{D}_S$. De Punder (2026) establishes conditions on ϕ under which $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi) > 0$ for an arbitrary strictly locally proper scoring rule S, provided that the step-size $\tilde{\alpha}$ is small enough. A model satisfying these conditions is called PRADA with respect to the local divergence space $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.

While De Punder (2026) leaves the estimation problem of the static deep parameters $\lambda \in \Lambda$ unspecified, the PRADA updates can be embedded directly into the quasi score-driven (QSD) framework of Blasques et al. (2023). Specifically, by setting the updating step as $\Delta\phi(y_t, \vartheta_{t|t-1}) \coloneqq \vartheta_{t|t} - \vartheta_{t|t-1} = \tilde{\alpha}\psi(y_t, \vartheta_{t|t-1})$, where $\tilde{\alpha} \in \mathbb{R}$ and $\psi : \mathcal{Y} \times \Theta \to \Theta$ is measurable with respect to $\mathcal{F}_t^{Y,Z} \coloneqq \sigma(\{Y_s, Z_s : s \leq t\}) \subseteq \mathcal{F}_t$ for some observable exogenous variables Z_t, we map the filter into their setting. Combined with the prediction step, this yields

$$\vartheta_{t+1|t} = \omega + \alpha\, \psi(y_t, Z_t, \vartheta_{t|t-1}, \lambda) + \beta\, \vartheta_{t|t-1}, \tag{4}$$

for real-valued parameters ω, α, and β. The restriction to $\mathcal{F}_t^{Y,Z} \subseteq \mathcal{F}_t$ makes estimation feasible while allowing the true measure to depend on a latent state, e.g., under a state-space specification.

The moment condition (2) and the recursion (4) together determine the filter entirely through ψ. The following example illustrates how classical time-series filters arise as special cases under quadratic loss.

Example 1. *For a location model ($k = 1$) under quadratic loss, the moment condition* (2) *reduces to* $\mathbb{E}_{\mathrm{P}_t}[Y_t - \vartheta_{t|t-1}] = 0$*, so that* $\psi(y, \vartheta) = y - \vartheta$*. Substituting into* (4) *yields*

$$\vartheta_{t+1|t} = \omega + \alpha y_t + (\beta - \alpha)\vartheta_{t|t-1},$$

corresponding to the ARMA(1,1) specification $y_t = \omega + \beta y_{t-1} + (\alpha - \beta)\varepsilon_{t-1} + \varepsilon_t$*.*

The second conditional moment is elicitable as the mean of Y_t^2 *(Fissler and Ziegel 2016), giving moment condition* $\mathbb{E}_{\mathrm{P}_t}[Y_t^2 - \vartheta_{t|t-1}] = 0$ *and hence* $\psi(y, \vartheta) = y^2 - \vartheta$*. For a volatility model ($k = 2$), with conditional mean zero, this implies the conditional variance prediction*

$$\vartheta_{t+1|t} = \omega + \alpha y_t^2 + (\beta - \alpha)\vartheta_{t|t-1},$$

which corresponds to the standard GARCH filter upon reparameterization.

The ARMA and GARCH filters above correspond to $\gamma = 2$ in the Barron loss. Section 2.2 replaces quadratic loss with the general Barron scoring function, which introduces data-adaptive robustness while retaining these filters as special cases.

2.2 BLADE updates

Barron (2019) proposes a rich class of loss functions nesting many well-known robust *location* estimators. The Barron loss (L_k^{B}) is infinitely differentiable in ϑ and its tuning

parameters $\gamma \in \bar{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty, \infty\}$ and $\xi > 0$, defining a Z-estimator through the gradient, $\psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi) \coloneqq \nabla_\vartheta S_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)$, where $S_k^{\mathrm{B}} \coloneqq -L_k^{\mathrm{B}}$ is the positively oriented scoring function with $S_k^{\mathrm{B}} : \mathcal{Y} \times \Theta \to \mathbb{R}_- \cup \{-\infty\}$, and $\mathbb{R}_- \coloneqq (-\infty, 0]$. Both the Barron scoring function and its gradient are given in Table (1) and illustrated in Figure 1.

Table 1: Barron score function and gradient for different values of γ.

	$S_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)$	$\psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)$
$\gamma = 2$	$-\frac{1}{2}\left(\frac{y^k - \vartheta}{\xi}\right)^2$	$\frac{y^k - \vartheta}{\xi^2}$
$\gamma = 0$	$-\log\left(1 + \frac{1}{2}\left(\frac{y^k - \vartheta}{\xi}\right)^2\right)$	$\frac{2(y^k - \vartheta)}{(y^k - \vartheta)^2 + 2\xi^2}$
$\gamma = -\infty$	$\exp\left(-\frac{1}{2}\left(\frac{y^k - \vartheta}{\xi}\right)^2\right) - 1$	$\frac{y^k - \vartheta}{\xi^2}\exp\left(-\frac{1}{2}\left(\frac{y^k - \vartheta}{\xi}\right)^2\right)$
otherwise	$-\frac{\lvert\gamma - 2\rvert}{\gamma}\left[\left(1 + \frac{(y^k - \vartheta)^2}{\xi^2\lvert\gamma - 2\rvert}\right)^{\gamma/2} - 1\right]$	$\frac{y^k - \vartheta}{\xi^2}\left(\frac{(y^k - \vartheta)^2}{\xi^2\lvert\gamma - 2\rvert} + 1\right)^{\frac{\gamma}{2} - 1}$

NOTE: This table reports the score function $S_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)$ introduced by Barron (2019) and its gradient $\psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)$, evaluated at four representative values of the robustness parameter γ. Here y denotes the observation, ϑ the time-varying parameter being updated, $\xi > 0$ a scale parameter, and k indexes the power of y entering the score. Specific values recover well-known loss functions: quadratic ($\gamma = 2$), Charbonnier ($\gamma = 1$), Cauchy ($\gamma = 0$), and Welsch ($\gamma \to -\infty$).

For any $\gamma \in \bar{\mathbb{R}}$ and $\xi \in \mathbb{R}_+$, the gradient of the Barron score implies the BLADE update

$$\phi_k^{\mathrm{B}}(y_t, \vartheta_{t|t-1}; \gamma, \xi) \coloneqq \vartheta_{t|t-1} + \alpha\psi_k^{\mathrm{B}}(y_t, \vartheta_{t|t-1}; \gamma, \xi). \tag{5}$$

Together with the prediction step, we obtain the following general formulation of the BLADE filter.

Definition 2. *The* BLADE($\bar{p}, \bar{q}$) filter *is given by*

$$\vartheta_{t+1|t} \coloneqq \omega + \sum_{i=1}^{\bar{p}} \alpha_i \psi_k^{\mathrm{B}}(y_{t-i+1}, \vartheta_{t-i+1|t-i}; \gamma, \xi) + \sum_{j=1}^{\bar{q}} \beta_j \vartheta_{t-j+1|t-j}, \tag{6}$$

where ψ_k^{B} *is given in Table* (1).

Generally, we will consider the BLADE(1,1) model and particularly the update step in

Equation (5), recovered by $\omega = 0$ and $\beta \equiv \beta_1 = 1$, when studying its theoretical properties. The latter case serves as the canonical Barron *update*.

Remark 1. *In the case that $\gamma = 2$, we fix $\xi = 1$, as otherwise an identification problem arises in* (6)*. This issue does not occur for the other BLADE filters, where $\gamma \neq 2$; in those cases ξ can therefore be estimated.*

In general, the choice of the scoring function determines the statistical functional targeted by the estimator (Gneiting 2011). The Barron scoring function S_k^{B} induces a family of location functionals indexed by γ, defined as

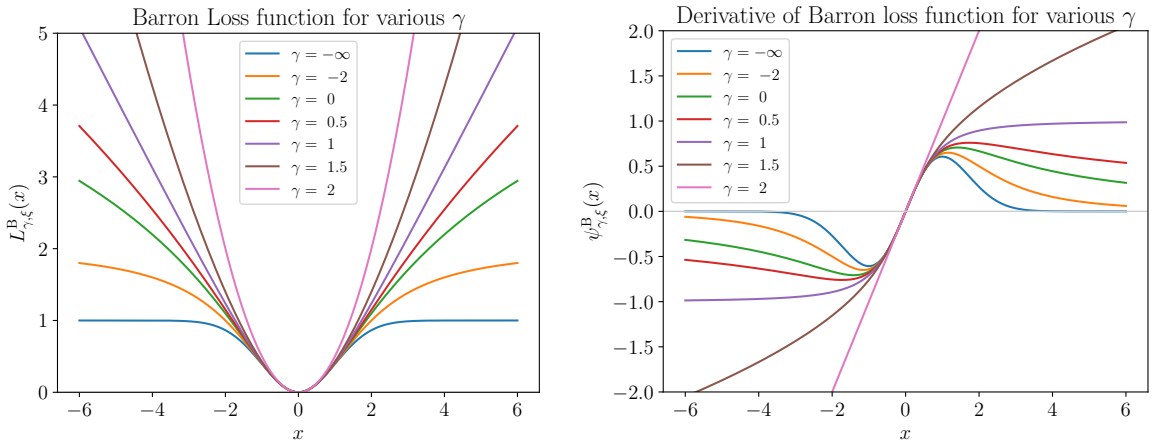
$$u^{\mathrm{B}}_{\gamma,\xi}(\mathrm{F}_{\vartheta^*}) \equiv \vartheta^* = \arg\max_{\vartheta \in \Theta} \mathbb{E}_{\mathrm{P}}\big[S_k^{\mathrm{B}}(Y, \vartheta; \gamma, \xi)\big]. \tag{7}$$

For the Barron score in Table (1) it is convenient to introduce $x_k \equiv x_k(y_t, \vartheta_{t|t-1}) \coloneqq y_t^k - \vartheta_{t|t-1}$, which represents the model residual, whose interpretation depends on the chosen state parametrization. For a conditional location model, the natural residual is $y_t - \vartheta_{t|t-1}$ (Blasques et al. 2023); for volatility models with mean zero it is common to use $y_t^2 - \vartheta_{t|t-1}$ (Andersen and Bollerslev 1998), so that $\vartheta_{t|t-1}$ tracks conditional variance and the residual reflects deviations in squared returns. More generally, any differentiable model-based residual can be used, provided it links $\vartheta_{t|t-1}$ to y_t in the chosen parametrization.

For each y, the Barron loss admits a unique minimizer in ϑ, so the Barron scoring function is *strictly consistent* for its own functional (7) (Gneiting 2011). Evaluated at x_1, strict consistency for the mean $u(\mathrm{P}_t) = \mathbb{E}_{\mathrm{P}_t}[Y_t]$ holds if and only if $\gamma = 2$; for $\gamma < 2$ the induced functional departs from the mean, with $\gamma = 1$ yielding a median-type and $\gamma \to -\infty$ a mode-type functional (Barron 2019) (see Appendix B.4). An analogous result holds for x_2, where strict consistency for the variance requires $\vartheta_{t|t-1} = \mathbb{E}_{\mathrm{P}_t}[Y_t^2]$ and $\mathbb{E}_{\mathrm{P}_t}[Y_t] = 0$. In many practical settings, such as under contamination or heavy-tailed errors, the robustness induced by smaller values of γ can result in improved filter updating, as the derivative

$\psi^{\mathrm{B}}_{\gamma,\xi}$ becomes bounded and redescending for $\gamma < 1$, preventing extreme observations from dominating the update step; see Figure 1.

Figure 1: The Barron Loss Function and its Gradient



(a) Barron loss function

(b) Derivative of Barron loss function

NOTE: The loss function (a) and its gradient (b) for different values of γ, the parameter ξ is fixed to 1. Note that the gradient shown here is equivalent to the formulation of $\psi^{\mathrm{B}}_{\gamma,\xi}(y_t, \vartheta_{t|t-1})$.

The implications of the choice of γ for the updating dynamics are illustrated in Example 2. For notational simplicity, write $\psi^{\mathrm{B}}_k(y, \vartheta; \gamma, \xi) \equiv \tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k(y, \vartheta))$, where

$$\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k) := \frac{x_k}{\xi^2}\left(1 + \frac{x_k^2}{|\gamma - 2|\,\xi^2}\right)^{\gamma/2-1}.$$

Example 2. *Consider the simple scenario in which the parameters are fixed to $\omega = 0$ and $\beta = 1$, so that the update of $\vartheta_{t|t-1}$ is given by*

$$\vartheta_{t+1|t} = \vartheta_{t|t-1} + \alpha\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}\big(x_k(y_t, \vartheta_{t|t-1})\big),$$

where, for $\gamma \notin \{-\infty, 0, 2\}$,

$$\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k(y, \vartheta)) = \frac{x_k}{\xi^2}\left(\frac{x_k^2}{\xi^2|\gamma - 2|} + 1\right)^{\gamma/2-1} = \frac{x_k}{\xi^2}W(\gamma), \qquad W(\gamma) := \left(1 + \frac{x_k^2}{\xi^2|\gamma - 2|}\right)^{\gamma/2-1}.$$

Since $1 + \frac{x_k^2}{\xi^2|\gamma-2|} \geq 1$, the scaling factor, $W(\gamma)$, satisfies $0 \leq W(\gamma) \leq 1$ for $\gamma \leq 2$, with

$W(\gamma) = 1$ *if* $\gamma = 2$. *Moreover, for fixed* $x_k(y_t, \vartheta_{t|t-1}) \neq 0$, $W(\gamma)$ *is decreasing in* γ *on* $(-\infty, 2)$*; this implies stronger down-weighting of large residuals for smaller values of* γ*, thereby increasing robustness against large observations.*

Notably, for $\gamma = 2$, $\xi = 1$ *(see Remark 1) and* $k = 1$*, the scaling disappears entirely and*

$$\psi^{\mathrm{B}}(y, \vartheta; \gamma, \xi) = y_t - \vartheta_{t|t-1},$$

which is proportional to the Gaussian SD update for a location model. Thus, for certain values of γ*, the proposed BLADE filter reduces to well-known SD formulations.*

Example 2 highlights two notable properties of the proposed BLADE filter. First, the filter nests well-known SD models as special cases. Second, the update equation can be reformulated to resemble an SD model with a normally distributed error term, but with an adaptive scaling function. Further analytical properties of the Barron loss function are provided in Appendix C.

3 Robustness

3.1 Contractive dominance

De Punder (2026) describes the conditions under which the update guarantees a reduction of the expected local divergence induced by the scoring rule S. A model satisfying these conditions is called PRADA. From a theoretical perspective, embedding the BLADE update within the class of PRADA updates is useful, as it directly links the updating mechanism to the reduction of a divergence measure. However, this property of an update does not quantify how the magnitude of the reduction varies across model specifications. The following theorem addresses this gap by deriving explicit conditions on the learning rate α under which the update with $\gamma_1 < \gamma_2$, i.e., a more robust update, achieves a larger expected

reduction in the local S-divergence while both updates remain PRADA. This formalizes the notion in which a more robust update dominates a less robust one in an expected divergence-reducing sense.

For this purpose, we propose the *expected divergence reduction difference* as a measure of relative update quality, allowing the updating distribution R_t to differ from the evaluation distribution P_t.

Definition 3. *Let $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ be a strictly locally proper scoring rule with induced local divergence $\mathbb{D}_S$ and consider two update rules $\phi_1, \phi_2 : \mathcal{Y} \times \Theta \to \Theta$. Let R_t be the distribution generating the updating draw $\check{Y}_t$. The expected divergence reduction difference is defined as*

$$\begin{aligned}\Delta\Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi_1 \| \phi_2) &\coloneqq \Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi_1) - \Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi_2) \\ &= \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{R}_t}\Big[S(\mathrm{F}_{\vartheta_{t|t}(\check{Y}_t, \gamma_1)}, Y_t)\Big] - \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{R}_t}\Big[S(\mathrm{F}_{\vartheta_{t|t}(\check{Y}_t, \gamma_2)}, Y_t)\Big],\end{aligned}$$

where $\vartheta_{t|t}(\check{Y}_t, \gamma_i) \coloneqq \phi_i(\check{Y}_t, \vartheta_{t|t-1})$.

Definition 3 measures the relative performance of two updates by the difference in their expected divergence reductions under a common updating measure R_t. A positive value of $\Delta\Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi_1 \| \phi_2)$ indicates that ϕ_1 contracts the expected divergence towards P_t more than ϕ_2, and a negative value reverses the ordering. This motivates the following dominance relation between update rules.

Definition 4. *Consider two update rules $\phi_1, \phi_2 : \mathcal{Y} \times \Theta \to \Theta$. Under the updating distribution R_t let $\Delta\Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi_1 \| \phi_2)$ be defined as in Definition 3. If $\Delta\Delta\mathbb{D}_S^{\mathrm{R}_t}(\phi_1 \| \phi_2) > 0$, we say that ϕ_1 P_t–contractively dominates ϕ_2 with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$.*

In line with the generalization error as the expected out-of-sample loss (Hastie et al. 2009, Chapter 7), we interpret $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1 \| \phi_2)$ as the difference in out-of-sample improve-

ment between the two updates when both updating and evaluation occur under P_t. Theorem 1 provides, under Assumption 1, explicit thresholds $\underline{\alpha}_t$ and $\bar{\alpha}_t$ such that for any $\alpha \in [\underline{\alpha}_t, \bar{\alpha}_t)$ the robust update with $\gamma_1 < \gamma_2 \leq 2$ P_t-contractively dominates the less robust update, while both updates remain PRADA. In particular, Assumption 1(iii) imposes expected concavity of the score with curvature bounded by c, analogous to Gorgi et al. (2024, Assumption 3), which controls the quadratic remainder and determines the upper threshold $\bar{\alpha}_t$. Assumption 1(v) is the expected gradient alignment condition of De Punder (2026, Definition 4), needed to establish the PRADA bound. For notational simplicity write $\psi_j(\check{Y}_t^k, \vartheta_{t|t-1}) \equiv \psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_j, \xi)$ for $j \in \{1, 2\}$.

Assumption 1.

(i) For all y, the map $\vartheta \mapsto S(\mathrm{F}_\vartheta, y)$ is twice continuously differentiable, and $\mathbb{E}_{\mathrm{P}_t}\left[\left|\frac{\partial S(\mathrm{F}_\vartheta, Y_t^k)}{\partial \vartheta}\Big|_{\vartheta=\vartheta_{t|t-1}}\right|\right] < \infty$.

(ii) For $\gamma_i \leq 2$ with $i \in \{1, 2\}$ and all $u \in [0, 1]$, $\vartheta_{t|t-1} + u\alpha\, \psi_i^{\mathrm{B}}(Y_t^k, \vartheta_{t|t-1}) \in \Theta$ holds P_t-a.s.

(iii) There exists $c \in (0, \infty)$ such that the expected Hessian satisfies $-c \leq \mathbb{E}_{\mathrm{P}_t}[H_S(Y_t, \vartheta)] < 0$ for all $\vartheta \in \Theta$.

(iv) For both $i \in \{1, 2\}$, $\mathbb{E}_{\mathrm{P}_t}\left[\psi_i^{\mathrm{B}}(Y_t^k, \vartheta_{t|t-1})^2\right] < \infty$.

(v) For both $i \in \{1, 2\}$, $\mathbb{E}_{\mathrm{P}_t}\left[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\right] \mathbb{E}_{\mathrm{P}_t}\left[\psi_i^{\mathrm{B}}(Y_t^k, \vartheta_{t|t-1})\right] > 0$.

Theorem 1. *Consider two BLADE updates, $\phi_1^{\mathrm{B}}, \phi_2^{\mathrm{B}}$, with $\gamma_1 < \gamma_2 \leq 2$ and let Y_t and $\check{Y}_t$ be independent copies with distribution P_t. Fix $\vartheta_{t|t-1} \in \Theta$ and suppose Assumption 1 holds. Then, $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) > 0$, for all $i \in \{1, 2\}$ and $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) \geq 0$, for any learning rate*

$\alpha \in [\underline{\alpha}_t, \bar{\alpha}_t)$, *where* $\underline{\alpha}_t := \max\{0, \underline{\alpha}_t^*\}$ *and*

$$\underline{\alpha}_t^* := \frac{\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\big(\psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}) - \psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\big)\Big]}{\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\,\psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2 - \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\,\psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\Big]},$$

$$\bar{\alpha}_t := \min_{i\in\{1,2\}} \frac{2}{c}\frac{\mathbb{E}_{\mathrm{P}_t}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big]\mathbb{E}_{\mathrm{P}_t}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big]}{\mathbb{E}_{\mathrm{P}_t}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\Big]},$$

where the integrated hessian term, $\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i))$*, is given by*

$$\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i)) := \int_0^1 (1-u)\, H_{\mathrm{S}}\Big(Y_t^k,\, \vartheta_{t|t-1} + u\alpha\, \psi_i^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big)\,\mathrm{d}u.$$

The lower threshold $\underline{\alpha}_t$ depends on two components. The numerator of $\underline{\alpha}_t^*$ measures the gap between the expected updates of the two filters, $\mathbb{E}_{\mathrm{P}_t}[\psi_2^{\mathrm{B}} - \psi_1^{\mathrm{B}}]$, weighted by the alignment with the expected score gradient $\mathbb{E}_{\mathrm{P}_t}[\nabla_\vartheta S]$. The denominator measures the gap in their Hessian-weighted second moments, which is non-negative under Assumption 1(iii). Heuristically, the denominator measures the extent to which the robust update reduces the second-moment contribution of observations relative to the less robust one. When this difference in second moments is large, large observations contribute heavily to $(\psi_2^{\mathrm{B}})^2$ but are attenuated in $(\psi_1^{\mathrm{B}})^2$, so the denominator is large and $\underline{\alpha}_t$ is small. Consequently, only a modest learning rate is required for the robust update to be P_t–contractively dominating over the less robust update.

The upper threshold $\bar{\alpha}_t$ is determined by the curvature constant c from Assumption 1(iii) and the first two moments of each ψ_i^{B} under P_t. It characterizes the largest learning rate at which both updates remain PRADA. Intuitively, $\alpha > \bar{\alpha}_t$ describes the regime in which the update 'overshoots' its target; although the update direction remains correct, the step size becomes too large to yield a reduction in divergence. Example C.1 illustrates both thresholds for Student's-t innovations and demonstrates that the interval $[\underline{\alpha}_t, \bar{\alpha}_t)$ is non-empty.

A complementary question in the score-driven literature concerns the choice of scaling matrix. Existing recommendations are typically derived from asymptotic filter error variance arguments under correct density specification (Beutner et al. 2026), leaving open the question of how to rank feasible scalings under a criterion that is well-defined irrespective of specification. Definition 4 also induces such a ranking through expected divergence reductions under Definition 3: rather than selecting scaling matrices heuristically, they can be compared directly in terms of their expected divergence reduction. Appendix E develops this connection in detail and Example 3 illustrates the comparison between scaling with the inverse Fisher scalar and the identity.

Example 3. *Consider a score-driven model with modeled conditional distribution* $Y_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0, \vartheta)$*, time-varying conditional variance* $\vartheta_{t|t-1}$*, and score* $s(y, \vartheta) = y^2/(2\vartheta^2) - 1/(2\vartheta)$ *with Fisher information* $\mathcal{I}_{t-1} = 1/(2\vartheta^2)$*. Compare two updates with scaling factors* $\mathcal{S}_{t-1}^{(1)} = \mathcal{I}_{t-1}^{-1} = 2\vartheta^2$ *(inverse Fisher) and* $\mathcal{S}_{t-1}^{(2)} = 1$ *(identity), under* $\mathrm{P}_t = \mathcal{N}(0, 2)$ *and* $\vartheta_{t|t-1} = 1$*. Since the Hessian* $H_{\mathrm{LogS}}(y, \vartheta) = 1/(2\vartheta^2) - y^2/\vartheta^3$ *depends on* y*, the quantities* $\Delta\Delta\mathbb{D}_{\mathrm{LogS}}^{\mathrm{P}_t}$ *and* $\Delta\mathbb{D}_{\mathrm{LogS}}^{\mathrm{P}_t}$ *are evaluated numerically as detailed in Appendix E.1, yielding*

$$\Delta\Delta\mathbb{D}_{\mathrm{LogS}}^{\mathrm{P}_t}(\phi_1\|\phi_2) > 0 \quad \textit{for all } \alpha \in (0,\, 0.109),$$

while both updates remain PRADA on the larger interval $(0,\, 0.195)$*. Intuitively, the inverse Fisher scaling doubles the effective step size relative to identity scaling; the threshold* $\alpha = 0.109$ *marks the point at which this amplification overshoots and the curvature penalty of the logarithmic scoring rule dominates.*

3.2 Contractive dominance under contamination

In this section, we examine the robustness of the BLADE update in a contaminated setting. We model contamination through the mixture measure (8),

$$\mathrm{P}_t^{\varepsilon} \coloneqq (1-\varepsilon)\mathrm{P}_t + \varepsilon \mathrm{H}_t, \qquad \varepsilon \in (0,1), \tag{8}$$

following the gross error model of Huber (1981, Eq. (4.4)), adapted to a conditional setting. Here P_t represents a clean data generating process and H_t a contamination, defined as the probability kernel $\mathrm{H}_t(\cdot) \coloneqq \mathbb{P}(V_t \in \cdot \mid \mathcal{F}_{t-1})$ for some auxiliary real-valued stochastic process $\{V_t\}_{t\in\mathbb{Z}}$ on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathcal{V} \subseteq \mathbb{R}$. Under Assumptions 1 and 2, Theorem 2 extends Theorem 1 to a contaminated updating distribution $\mathrm{P}_t^{\varepsilon}$, providing explicit lower and upper bounds, $\underline{\varepsilon}$ and $\bar{\varepsilon}$, on the contamination level such that for every $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon})$ the robust BLADE update with $\gamma_1 < \gamma_2$, P_t–contractively dominates the less robust update while both updates remain PRADA.

Assumption 2.

(i) For $\gamma \leq 2$ and all $u \in [0,1]$, $\vartheta_{t|t-1} + u\alpha\, \psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma, \xi) \in \Theta$ holds $\mathrm{P}_t^{\varepsilon}$-a.s.

(ii) For both $i \in \{1,2\}$, $\mathbb{E}_{\mathrm{H}_t}\big[\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)^2\big] < \infty$.

Theorem 2. *Consider two BLADE updates, $\phi_1^{\mathrm{B}}, \phi_2^{\mathrm{B}}$, with $\gamma_1 < \gamma_2 \leq 2$, and let $Y_t^k \sim \mathrm{P}_t$ and $\check{Y}_t^k \sim \mathrm{P}_t^{\varepsilon}$ be independent, where $\mathrm{P}_t^{\varepsilon}$ (8) denotes a contaminated updating distribution with arbitrary contaminating distribution H_t. Suppose Assumption 1 and Assumption 2 hold, furthermore suppose that $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) \geq 0$ and $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) > 0$ for all $i \in \{1,2\}$. Then $\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_i^{\mathrm{B}}) > 0$ for all $i \in \{1,2\}$ and $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) \geq 0$ for any $\varepsilon \in (\underline{\varepsilon}, \bar{\varepsilon})$, where $\bar{\varepsilon} \coloneqq \min_{i\in\{1,2\}} \bar{\varepsilon}(\gamma_i, \alpha)$ and*

$$\underline{\varepsilon} \equiv \underline{\varepsilon}(\gamma_1, \gamma_2, \alpha) \coloneqq \max\left\{0, \frac{-\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}})}{\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) - \Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}})}\right\},$$
$$\bar{\varepsilon}(\gamma_i, \alpha) \coloneqq \min\left\{\frac{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}})}{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) - \Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}})}, 1\right\}.$$

The key assumption, $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \geq 0$, requires that the dominance ordering already holds when updating entirely under H_t. Corollary 1 provides a sufficient condition for this assumption in terms of the tail behavior of H_t. For notational simplicity define the conditional measure; for any $M > 0$, with $\mathrm{H}_t(|\check{x}_k| > M) > 0$, $\mathrm{H}_t^{>M} \coloneqq \mathrm{H}_t(\cdot \mid |\check{x}_k| > M)$.

Corollary 1. *Under the setting of Theorem 2, let $\check{x}_k = x_k(\check{Y}_t^k, \vartheta_{t|t-1})$ be the residual map evaluated in $\check{Y}_t^k$ and suppose there exists $M > 0$ such that $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{>M}}(\phi_1||\phi_2) > 0$ and*

$$\underline{\mathrm{H}}_t(|\check{x}_k| > M) \;\geq\; \max\left\{0,\; \frac{-\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1\|\phi_2)}{\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{>M}}(\phi_1\|\phi_2) - \Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1\|\phi_2)}\right\},$$

where the right-hand side lies in $[0,1]$. Then $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \geq 0$ and $\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_i^{\mathrm{B}}) > 0$ for all $i \in \{1,2\}$, for every $\varepsilon \in (\underline{\varepsilon}, \bar{\varepsilon})$, where $\underline{\varepsilon}$ and $\bar{\varepsilon}$ are as defined in Theorem 2.

The condition $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{>M}}(\phi_1||\phi_2) > 0$ states that, conditional on the contamination draw lying sufficiently far from the current state $\vartheta_{t|t-1}$, the robust update yields a larger expected divergence reduction. This is the natural regime: for larger $|\check{x}_k|$, the more robust Barron score attenuates the update relative to the less robust specification, so the second-order variance reduction dominates the first-order loss. In the scenario where $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1||\phi_2) \geq 0$, the robust update is also favored for moderate contamination draws, and the $\max\{0,\cdot\}$ structure in Corollary 1 yields $\underline{\varepsilon} = 0$, so no tail-mass requirement is needed. However, when $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1||\phi_2) < 0$, moderate draws penalize the robust update. In such cases, the tail-mass bound quantifies the amount of probability H_t must place on extreme observations to compensate for the tail benefit.

The relation between $\alpha, \gamma_1, \gamma_2, \underline{\varepsilon}$ and $\bar{\varepsilon}$ is characterized by Theorem 2. To provide some intuition, consider updates $\vartheta_{t|t}$ evaluated under the quadratic scoring rule $S(x,y) = -(x-y)^2$, with non-robust update $\gamma_2 = 2$ and robust update $\gamma_1 < 2$, and updates drawn from (8) where H_t is Student's-t with 4 degrees of freedom and scale 3.

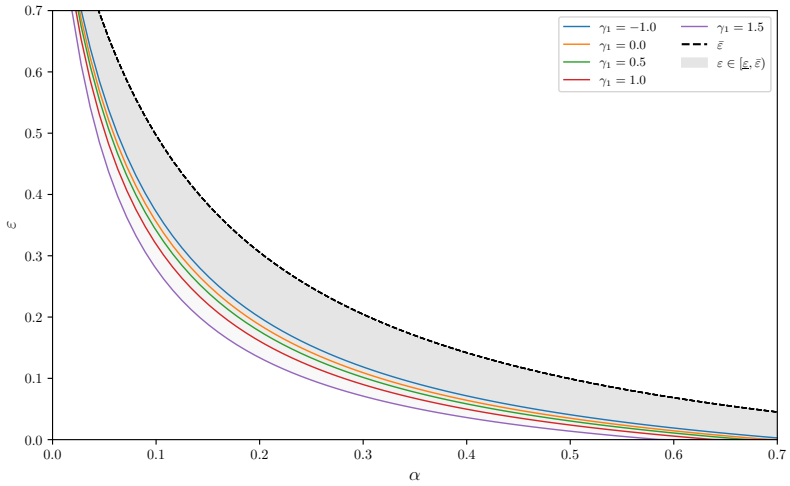
Figure 2 displays $\underline{\varepsilon}(\gamma_1, \gamma_2, \alpha)$ and $\bar{\varepsilon}(\alpha) = \min_{i\in\{1,2\}} \bar{\varepsilon}(\gamma_i, \alpha)$ for several values of γ_i,

with the shaded region indicating the valid contamination interval $[\underline{\varepsilon}, \bar{\varepsilon})$. The curves are convex and decreasing in α: at small α the update step is small and a large contamination fraction is needed to guarantee improvement, while at large α the threshold is already small and further increases in α yield little additional reduction. The curves shift upward for smaller γ_1 over the range shown, indicating that within this calibration a more robust update requires a larger contamination fraction to dominate the non-robust benchmark. The upper bound $\bar{\varepsilon}$ is determined by $\bar{\varepsilon}(\gamma_2, \alpha)$ for every value of γ_1 shown, reflecting that the non-robust update $\gamma_2 = 2$ is the first to violate the PRADA condition as contamination increases. Importantly, $\underline{\varepsilon}$ captures only whether γ_1 dominates $\gamma_2 = 2$, not the magnitude of the improvement; a smaller γ_1 requires a larger contamination fraction to dominate but yields a substantially larger MSE reduction once it does. Example C.2 furthermore numerically illustrates a vertical slice of Figure 2 at $\alpha = 0.5$ and $\gamma_1 = 1$.

The results of the previous two subsections provide sufficient conditions under which a more robust update (ϕ_1^{B}) P_t-contractively dominates a less robust update (ϕ_2^{B}). In practice the underlying properties of the DGP are unknown, so the optimal degree of robustness must be determined from the data. The continuous parameterization of the Barron loss enables data-driven selection of γ_1 by minimizing an out-of-sample criterion over a grid. Section 5 implements this and shows that γ_1^* departs systematically from the quadratic case as contamination increases.

Related notions of robustness appear in the literature. Calvet et al. (2015) study robustness in a state-space framework by requiring that contamination induces only bounded distortions of the filtered state. In contrast to Definition 3, the robustness condition is pointwise in the state and observations, and does not rely on an independent evaluation draw (see Appendix C.7 for more details). A different complementary viewpoint is provided by Hampel et al. (1986), who characterize robustness through the first-order influence of infinitesimal contamination on the estimator; this perspective is taken up in the next sub-

Figure 2: The relation between $\underline{\varepsilon}$, $\bar{\varepsilon}$ and α with $\mathrm{H}_t \sim t_4(0, 3^2)$



NOTE: Contamination threshold $\underline{\varepsilon}(\gamma_1, \gamma_2, \alpha)$ (solid) and PRADA upper bound $\bar{\varepsilon}(\gamma_i, \alpha)$ (dashed) as a function of the learning rate α for several values of γ_i, with $\gamma_2 = 2$ and $\check{Y}_t \sim \mathrm{H}_t$ Student's-$t(4)$ with scale 3. For a given (γ_1, α), any contamination fraction $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon})$ guarantees that the robust update outperforms the non-robust update in expected MSE and both updates remain PRADA under $\mathrm{P}_t^{\varepsilon}$.

section in the context of a BLADE update.

3.3 B-robustness

Theorem 2 characterizes robustness of the expected divergence reduction difference. A complementary perspective due to Hampel et al. (1986) considers the first-order effect of an infinitesimal contamination on the estimator. This notion, known as *B-robustness*, requires that such an infinitesimal perturbation cannot have an arbitrarily large effect, and is fully characterized by the boundedness of the influence function, introduced below.

Following Hampel et al. (1986) and Hampel (1974), we model a small contamination of P_t by (8). The corresponding perturbed parameter is $\vartheta_\varepsilon := u(\mathrm{P}_t^{\varepsilon})$, with $u : \mathcal{P} \to \Theta$. The corresponding influence function measures the first–order effect of such a contamination on the value of the functional.

Formally, the influence function is defined as the Gâteaux derivative of u at P_t in the

direction of a point mass δ_y, that is,

$$\mathrm{IF}(y; u, \mathrm{P}_t) \coloneqq \frac{\mathrm{d}}{\mathrm{d}\varepsilon} u\big((1-\varepsilon)\mathrm{P}_t + \varepsilon\delta_y\big)\Big|_{\varepsilon=0^+} = \lim_{\varepsilon\to 0^+} \frac{u\big((1-\varepsilon)\mathrm{P}_t + \varepsilon\delta_y\big) - u(\mathrm{P}_t)}{\varepsilon}, \tag{9}$$

for $y \in \mathcal{Y}$. Intuitively, $\mathrm{IF}(y; u, \mathrm{P}_t)$ is the infinitesimal change in ϑ when an observation at y is given an infinitesimally larger weight.

The relevance of the influence function for robustness is summarized by the gross–error sensitivity

$$\zeta^*(u, \mathrm{P}_t) \coloneqq \sup_{y\in\mathcal{Y}} |\mathrm{IF}(y; u, \mathrm{P}_t)|,$$

which quantifies the maximum effect that an infinitesimal contamination at any point y can exert on the estimator. If $\zeta^*(u, \mathrm{P}_t) < \infty$, the estimator is said to be *B–robust* at P_t, since a small amount of contamination cannot produce an unbounded asymptotic effect. By M–estimator theory (see, e.g., Hampel et al. 1986, Eq. (2.3.5)), the influence function of the Barron functional (7) at P_t is

$$\mathrm{IF}(y; u^{\mathrm{B}}_{\gamma,\xi}, \mathrm{P}_t) = M_\gamma(\mathrm{P}_t)^{-1}\, \psi^{\mathrm{B}}_k(y, u^{\mathrm{B}}_{\gamma,\xi}(\mathrm{P}_t); \gamma, \xi), \tag{10}$$

where $M_\gamma(\mathrm{P}_t) \coloneqq -\mathbb{E}_{\mathrm{P}_t}\left[\nabla_\vartheta \psi^{\mathrm{B}}_k(Y, u^{\mathrm{B}}_{\gamma,\xi}(\mathrm{P}_t); \gamma, \xi)\right]$ is the scalar sensitivity term; see Appendix C.4 for more details.

For the Barron functional, the robustness properties depend entirely on the tail behavior of the score ψ^{B}_k. The following proposition formalizes this relationship and identifies the parameter region for which B–robustness holds.

Proposition 1. *Let P^ε_t be the contaminated measure defined in* (8)*, and let $u^{\mathrm{B}}_{\gamma,\xi}(\mathrm{P}_t)$ be the Barron functional, defined in* (7) *with corresponding influence function IF$(y; u^{\mathrm{B}}_{\gamma,\xi}, \mathrm{P}_t)$* (10)*. The Barron functional, is B–robust at P_t for all ξ if and only if $\gamma \le 1$.*

The tuning parameter γ thus acts as a continuous robustness index, with $\gamma \le 1$ guar-

anteeing a bounded influence function and more negative values of γ providing stronger protection against outlying observations. Furthermore, the robustness discussion can also be related to asymptotic efficiency; this is discussed in Appendix C.6.

4 Estimation

This section establishes the consistency and asymptotic normality of the quasi-likelihood estimator (QLE) $\hat{\lambda}_T$ of the static parameter vector $\lambda \in \Lambda$ of the BLADE filter $\vartheta_{t|t-1}(\lambda)$. We assume correct specification for some fixed $\xi \in (0, \infty)$ and shape parameters $\gamma_{\text{fil}}, \gamma_{\text{mom}} \leq 2$ governing the filter update and the estimation moment condition, respectively. Specifically, following Blasques et al. (2023), the true conditional distribution is assumed parametric in the true time-varying parameter $\vartheta_t \equiv \vartheta_{t|t-1}(\lambda_0)$ generated under $\lambda_0 \in \Lambda$, such that $\tilde{\mathrm{P}}_t \equiv \tilde{\mathrm{P}}_{\vartheta_t}$. Formally, the conditional distribution $\tilde{\mathrm{P}}_t$ is defined as the probability kernel $\tilde{\mathrm{P}}_t(\cdot) \coloneqq \mathbb{P}(Y_t \in \cdot \mid \mathcal{F}_{t-1}^{Y,Z})$, on $(\mathcal{Y}, \mathcal{G})$.

For the Barron loss, the conditional moment condition of Theorem 1 in Blasques et al. (2023) then generalizes to

$$\mathbb{E}_{\tilde{\mathrm{P}}_t}\big[\psi^{\mathrm{B}}_{\gamma_{\text{mom}},\xi}(Y_t^k - \vartheta_t)\big] = 0, \qquad t \in \mathbb{Z}, \tag{11}$$

reducing to their specification under the identity $\psi^{\mathrm{B}}_{\gamma_{\text{mom}},\xi}$. The shape parameter γ_{mom} identifies ϑ_t as the corresponding functional of $\tilde{\mathrm{P}}_t$, while γ_{fil} governs the generating filter recursion. Although the two need not coincide, we set $\gamma_{\text{fil}} = \gamma_{\text{mom}} =: \gamma$, so that the local divergence reduced by the filter at each update step coincides with the divergence minimized in estimation.

This framework allows estimation without specifying the parametric form of $\tilde{\mathrm{P}}_t$ (Durbin 1960; Godambe 1960). Formally, define the estimating function, $h_t \equiv h_t^k(\lambda) \coloneqq \psi^{\mathrm{B}}_{\gamma,\xi}\big(Y_t^k - \vartheta_{t|t-1}(\lambda)\big) \in \mathbb{R}$, depending on Y_t and $\vartheta_{t|t-1} \equiv \vartheta_{t|t-1}(\lambda)$. By the moment condition (11)

and correct specification, the estimating function is conditionally unbiased at λ_0, i.e., $\mathbb{E}_{\tilde{\mathrm{P}}_t}\big[h_t(\lambda_0)\big] = \mathbb{E}_{\tilde{\mathrm{P}}_t}\big[\psi^{\mathrm{B}}_{\gamma,\xi}\big(Y_t^k - \vartheta_{t|t-1}(\lambda_0)\big)\big] = 0$ for all $t \in \mathbb{Z}$. Since $\vartheta_{t|t-1}(\lambda)$ depends on the unobserved pre-sample $\{y_t : t \leq 0\}$, in practice it is replaced by the feasible recursion $\hat{\vartheta}_{t|t-1}(\lambda)$ run from $t \geq 1$ with arbitrary initialization $\hat{\vartheta}_{1|0}$ obtaining

$$\hat{\vartheta}_{t+1|t}(\lambda) = \omega + \alpha\psi^{\mathrm{B}}_{\gamma,\xi}(y_t, \hat{\vartheta}_{t|t-1}(\lambda)) + \beta\hat{\vartheta}_{t|t-1}(\lambda), \qquad \text{for } t \geq 1.$$

The estimator solves the estimating equation $\hat{\lambda}_T = \arg\min_{\lambda\in\Lambda} \|\hat{\mathrm{G}}_T(\lambda)\|$, where the infeasible and feasible estimation vector components are given by $g_t \equiv g_t^k(\lambda) \coloneqq \frac{h_t(\lambda)}{\sigma_t^2(\lambda)} \cdot \frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}$ and $\hat{g}_t \equiv \hat{g}_t^k(\lambda) \coloneqq \frac{\hat{h}_t(\lambda)}{\hat{\sigma}_t^2(\lambda)} \cdot \frac{\partial\hat{\vartheta}_{t|t-1}(\lambda)}{\partial\lambda}$ respectively. Similarly, define $\mathrm{G}_T(\lambda) \coloneqq \frac{1}{T}\sum_{t=1}^T g_t(\lambda)$ and $\hat{\mathrm{G}}_T(\lambda) \coloneqq \frac{1}{T}\sum_{t=1}^T \hat{g}_t(\lambda)$.

The estimating equation $\hat{G}_T(\lambda)$ contains the functions $\hat{h}_t(\lambda), \frac{\partial\hat{\vartheta}_{t|t-1}(\lambda)}{\partial\lambda}, \hat{\sigma}_t^2(\lambda)$. These functions are approximations to the functions $h_t(\lambda), \frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}, \sigma_t^2(\lambda)$, which depend on the unobserved pre-sample. An approximation for $\sigma_t^2 = \mathbb{E}_{\tilde{\mathrm{P}}_t}[h_t^2(\lambda)]$ is recursively computed from $\{y_1, \ldots, y_T\}$. Blasques et al. (2023) note that $\hat{\sigma}^2_{t|t-1}$ may be chosen proportional to either $\hat{\vartheta}_{t|t-1}$ or $\hat{\vartheta}^2_{t|t-1}$, depending on the order of the conditional moment. The specific scaling of $\hat{\sigma}^2_{t|t-1}$ is without loss of generality and amounts only to a reparameterization of the estimating equation. Furthermore, denote the marginal stationary distribution as $\bar{\mathrm{P}} \coloneqq \mathrm{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathrm{P}_t$.

Under the following sufficient conditions the estimator $\hat{\lambda}$ is consistent and asymptotically normal.

Assumption 3.

(i) Let $(Y_t, \mathcal{F}_t^{Y,Z})_{t\in\mathbb{Z}}$ *be strictly stationary and ergodic. Moreover let* $\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+|Y_t|\right] < \infty$.

(ii) Let $\Lambda \subseteq \mathbb{R}^p$ *be compact,* $\gamma \leq 2$ *and let the updating equation* (4) *satisfy,* $\frac{|\alpha|}{\xi^2} + |\beta| < 1$.

(iii) Assume that there exists a constant $\underline{\sigma}^2 > 0$ *such that* $\inf_{\lambda\in\Lambda} \sigma_t^2(\lambda) \geq \underline{\sigma}^2$ *almost surely*

and $\sup_{\lambda\in\Lambda}|\hat{\sigma}_t^2(\lambda)-\sigma_t^2(\lambda)| \xrightarrow{a.s.} 0$ *as* $t\to\infty$.

(iv) Assume that for each t*, the maps* $\lambda\mapsto h_t(\lambda)$*,* $\lambda\mapsto\sigma_t^2(\lambda)$ *and* $\lambda\mapsto\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}$ *are continuous on* Λ*, and the population moment condition identifies* λ_0*, that is, the condition* $\mathbb{E}_{\bar{\mathrm{P}}}[g_t(\lambda_0)]=0$ *is satisfied if and only if* $\lambda=\lambda_0$.

(v) Let

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\sup_{\lambda\in\Lambda}|\vartheta_{t|t-1}(\lambda)|\right]<\infty,\quad \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\sup_{\lambda\in\Lambda}\left\|\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\right\|\right]<\infty.$$

(vi) Let

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\sup_{\lambda\in\Lambda}\left|\frac{h_t(\lambda)}{\sigma_t(\lambda)}\right|^2\right]<\infty,\qquad \mathbb{E}_{\bar{\mathrm{P}}}\left[\sup_{\lambda\in\Lambda}\left\|\frac{1}{\sigma_t(\lambda)}\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\right\|^2\right]<\infty.$$

Theorems 3 and 4 state the main asymptotic results on the estimated parameters.

Theorem 3. *Fix* $\gamma\le 2$*,* $\xi\in(0,\infty)$ *and* $k>0$*. Let* $\{Y_t\}_{t\in\mathbb{Z}}$ *be generated by the model and satisfy the conditional moment restrictions* $\mathbb{E}_{\tilde{\mathrm{P}}_t}\big[\psi^{\mathrm{B}}_{\gamma,\xi}\big(Y_t^k-\vartheta_{t|t-1}(\lambda_0)\big)\big]=0$*,* $\forall t\in\mathbb{Z}$*, while* $\{\vartheta_{t|t-1}\}_{t\in\mathbb{Z}}$ *satisfies the BLADE filter recursion* (6)*. Suppose Assumption 3 holds and let* $\hat{\lambda}_T$ *be any sequence satisfying the estimating equation*

$$\hat{G}_T(\hat{\lambda}_T)=0.$$

Then

$$\hat{\lambda}_T \xrightarrow{a.s.} \lambda_0,\qquad \text{as } T\to\infty.$$

Define the Jacobian

$$\mathcal{J}_0 \coloneqq -\mathbb{E}_{\bar{\mathrm{P}}}\left[\frac{\partial}{\partial\lambda^\top}g_t(\lambda)\bigg|_{\lambda=\lambda_0}\right].$$

Under the following additional sufficient conditions asymptotic normality of the estimated parameters is established.

Assumption 4.

(i) $\lambda_0 \in \mathrm{Int}(\Lambda)$.

(ii) *The moment condition* $\mathbb{E}_{\tilde{\mathrm{P}}_t}[|Y_t|^{2k+\delta}] < \infty$ *is satisfied for some* $\delta > 0$,

(iii) *For some neighbourhood* $V(\lambda_0)$ *of* λ_0.

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\sup_{\lambda\in V(\lambda_0)}\left\|\frac{1}{\sigma_t^3(\lambda)}\,\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\,\frac{\partial\sigma_t^2(\lambda)}{\partial\lambda^\top}\right\|^2\right] < \infty, \qquad \mathbb{E}_{\bar{\mathrm{P}}}\left[\sup_{\lambda\in V(\lambda_0)}\left\|\frac{1}{\sigma_t(\lambda)}\,\frac{\partial^2\vartheta_{t|t-1}(\lambda)}{\partial\lambda\partial\lambda^\top}(\lambda)\right\|^2\right] < \infty.$$

(iv) $\varrho^{-t}\sup_{\lambda\in\Lambda}\left\|\frac{\partial\hat{\sigma}_t^2(\lambda)}{\partial\lambda} - \frac{\partial\sigma_t^2(\lambda)}{\partial\lambda}\right\| \to 0$ *a.s. as* $t\to\infty$.

(v) *The Jacobian* $\mathcal{J}_0$ *exists and is non–singular, and* $\mathbb{E}_{\bar{\mathrm{P}}}[g_t(\lambda_0)g_t(\lambda_0)^\top] < \infty$.

Theorem 4. *Let Assumptions 3 and 4 hold. Then*

$$\sqrt{T}\,(\hat{\lambda}_T - \lambda_0) \;\xrightarrow{d}\; \mathcal{N}\Big(0,\; \mathcal{J}_0^{-1}\Omega_0\mathcal{J}_0^{-1\top}\Big),$$

where $\Omega_0 \coloneqq \mathbb{E}_{\bar{\mathrm{P}}}[g_t(\lambda_0)g_t(\lambda_0)^\top]$.

Remark 2. *Theorems 3 and 4 presume the filter* (6) *is correctly specified. When this fails, the estimator converges not to a true parameter but to a pseudo-true value* λ^*. *One may either invoke Blasques et al. (2023, Theorem 2) directly, characterizing* λ^* *as a Kullback–Leibler projection, or extend that result by replacing the logarithmic scoring rule with a Barron scoring function* S *that is strictly consistent for* ϑ_t, *giving the functional-level analogue of that projection, to which it reduces when* S *is the log-score.*

5 Monte Carlo

5.1 Simulation design

We conduct a Monte Carlo study to assess the performance and evaluate the robustness properties of the BLADE filter under non-contaminated and contaminated training envi-

ronments. We consider a volatility model of the form

$$Y_t = \sqrt{\vartheta_{t|t-1}}\,\eta_t, \qquad \eta_t \overset{i.i.d.}{\sim} t_{10},$$
$$\vartheta_{t+1|t} = \omega + \alpha\big(Y_t^2 - \vartheta_{t|t-1}\big) + \beta\vartheta_{t|t-1}.$$

This DGP corresponds to a standard GARCH(1,1) specification and coincides with a BLADE filter based on the quadratic Barron loss, i.e., $\psi_2^{\mathrm{B}}(y_t, \vartheta_{t|t-1}; \gamma = 2, \xi = 1)$. The innovations (η_t) of the DGP are drawn from a Student–t distribution with 10 degrees of freedom. The true parameters are set to $(\omega_0, \alpha_0, \beta_0) = (0.02, 0.10, 0.88)$, which imply a persistent volatility process comparable to daily equity index returns. The recursion is initialized at the unconditional variance, $\vartheta_{0|-1} = \frac{\omega_0}{1-\alpha_0-\beta_0}$.

The simulated sample consists of 2500 observations for estimation and 1000 observations for evaluation. Contamination is introduced *exclusively* in the training sample, while the test sample remains non-contaminated. In line with the contaminated measure in (8), the observations are drawn from a mixture of a true conditional distribution P_t and a contamination measure H_t, where H_t denotes the conditional distribution induced by adding a heavy tailed shock with scale proportional to $\sqrt{\vartheta_{t|t-1}}$ to a clean draw from P_t. Operationally, the non-contaminated sequence is first generated and observational outliers are subsequently added to the training set according to, $y_t^{\mathrm{obs}} = y_t^{\mathrm{nc}} + I_t u_t$, where $I_t \sim$ Bernoulli(ε) determines whether an observation is contaminated and $\varepsilon \in \{0.00, 0.01, 0.02, 0.03\}$ denotes the fraction of contaminated observations.

The contamination component is specified as

$$u_t = \kappa\sqrt{\vartheta_{t|t-1}}\,Z_t, \qquad Z_t \sim t_3,$$

so that outliers scale proportionally with the prevailing conditional volatility. We set $\kappa = 10$,

generating large volatility proportional shocks. This design isolates the effect of contaminated parameter estimation on subsequent out-of-sample forecasting performance.

Remark 3. *Theorems 3 and 4 require the BLADE filter to be correctly specified for the target functional, which for the DGP above holds at $\gamma_{fil} = 2$. By design the Monte Carlo departs from this benchmark, taking $\gamma_{fil} < 2$ to increase robustness, so the resulting estimates are pseudo-true values rather than the λ_0 of the asymptotic theory. This is deliberate: the experiment is meant to show that under contamination the most useful filters are precisely the robust ones that fall outside the correctly-specified case.*

For each generated time series, BLADE volatility models are estimated with ξ fixed at 1 and robustness parameter γ_{fil} varying over a grid using QLE developed in Section 4. For each specification, we construct one-step-ahead forecasts $\hat{\vartheta}_{t|t-1}$ and evaluate forecast errors relative to the latent conditional variance path, $\delta_t = \vartheta^{\text{true}}_{t|t-1} - \hat{\vartheta}_{t|t-1}$. We report the mean absolute error (MAE) and the mean squared error (MSE), the latter of which corresponds to a Barron loss with $\gamma = 2$ and $\xi = 1$. We additionally consider the Barron loss at $\gamma = 1.5$ and $\gamma = 4$ for evaluation. The optimal robustness parameter γ^*_{fil} is selected separately for each evaluation loss.

As a benchmark, we include the β_t–GARCH(1,1) model of Harvey and Chakravarty (2008), defined by

$$\vartheta_{t+1|t} = \omega + \alpha \frac{\nu + 1}{\nu - 2 + \eta^2_{t|t-1}} \eta^2_{t|t-1} \vartheta_{t|t-1} + \beta \vartheta_{t|t-1},$$

where $\eta_t = Y_t / \sqrt{\vartheta_{t|t-1}}$, and parameters are estimated by maximum likelihood. All results are averaged across 1000 Monte Carlo replications.

5.2 Results

Table 2 reports the loss minimizing robustness parameter γ^*_{fil} across contamination levels. Under non-contaminated data ($\varepsilon = 0$), the optimal γ_{fil} slightly exceeds two for all loss criteria, indicating that little robustness is required when the model is correctly specified. In this setting, efficiency can be fully exploited and a sensitive update yields the lowest forecast error. As contamination increases, the loss-minimizing shape parameter γ^*_{fil} departs systematically from the quadratic case $\gamma_{\text{fil}} = 2$, which corresponds to the true data-generating process and continues to govern the evaluation sample. Despite the fact that performance is assessed under the original quadratic DGP, updates using $\gamma_{\text{fil}} < 2$ become optimal when estimation is conducted on contaminated data. For $\varepsilon = 0.03$, the optimal value for γ_{fil} is approximately 1.3 under MAE and MSE and remains below 1.7 even when evaluated under the more tail-sensitive Barron–4 loss. In line with Section 3.1, this shift reflects that, under contaminated estimation environments, the use of the quadratic loss ($\gamma = 2$) in the update step, is no longer optimal in terms of generalization error. Instead, values $\gamma_{\text{fil}} < 2$, which induce functionals that are less sensitive to extreme observations, yield a larger reduction in out-of-sample loss.

Table 2: Optimal γ^*_{fil} across contamination levels.

	$\varepsilon = 0.00$	$\varepsilon = 0.01$	$\varepsilon = 0.02$	$\varepsilon = 0.03$
MAE	2.206	1.556	1.455	1.333
Barron 1.5	2.141	1.677	1.455	1.333
MSE	2.108	1.677	1.455	1.333
Barron 4	2.043	1.798	1.697	1.636

NOTE: Optimal values of γ^*_{fil} selected for each evaluation criterion and contamination level ε. For MAE, Barron losses, and MSE, γ^*_{fil} minimizes the average out-of-sample loss. The models are trained on a sample of 2500 observations and evaluated on 1000 observations. Results are based on Monte Carlo averages over 1000 Monte Carlo replications.

Table 3 presents the corresponding optimal loss values and compares them to the β_t–GARCH(1,1) model. Forecast error increases with contamination under all criteria. For

example, the optimal MSE rises from 0.0049 in the non-contaminated setting to 0.0147 when $\varepsilon = 0.03$. Similar patterns are observed for MAE and Barron losses. While contamination mechanically increases forecast error, the increase is substantially mitigated when γ_{fil} is lowered to induce more robust updates. The lower panel of Table 3 reports a performance comparison between the β_t–GARCH(1,1) benchmark and BLADE. Under non-contaminated data, differences are small, indicating comparable performance. However, as contamination increases, the gap widens substantially across all evaluation criteria. This suggests that, while the heavy-tailed score specification of β_t–GARCH provides some robustness, training in a contaminated setting still distorts parameter estimates and deteriorates out-of-sample performance.

Table 3: Optimal loss values

	$\varepsilon = 0.00$	$\varepsilon = 0.01$	$\varepsilon = 0.02$	$\varepsilon = 0.03$
		BLADE		
MAE	5.14	6.81	6.61	6.33
Barron 1.5	0.22	0.53	0.59	0.60
MSE	0.49	1.22	1.42	1.47
Barron 4	0.37	1.53	2.71	3.57
		Difference (β_t-GARCH − BLADE)		
MAE	0.47	7.56	17.11	27.13
Barron 1.5	0.04	0.83	3.16	8.92
MSE	0.05	1.61	6.64	22.04
Barron 4	0.00	0.24	2.38	22.98

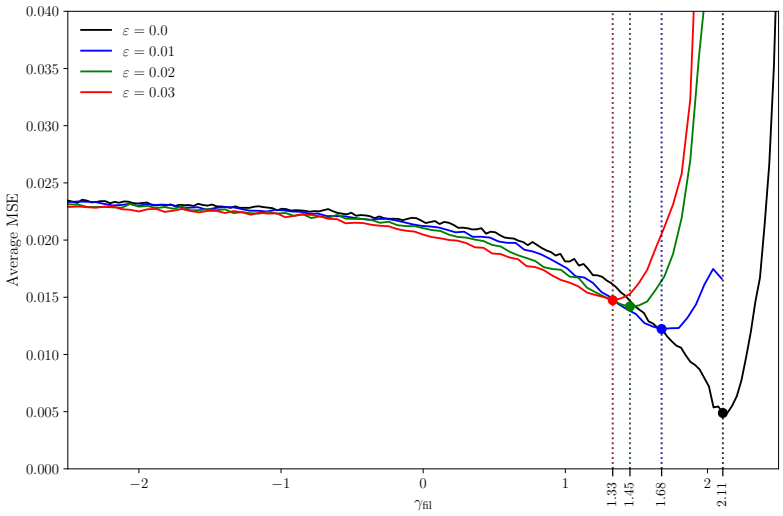
NOTE: Upper panel: BLADE losses scaled by a factor of 100 at the optimal γ_{fil}. Lower panel: β_t-GARCH(1,1) minus BLADE scaled by a factor of 100. Positive values indicate BLADE improvements. The models are trained on a sample of 2500 observations and evaluated on 1000 observations. Reported values are Monte Carlo averages over 1000 replications.

Figure 3 illustrates these findings for the MSE criterion by plotting the full loss surface as a function of γ_{fil}. In the non-contaminated case, the loss is minimized at values of γ_{fil} close to the quadratic specification $\gamma_{\text{fil}} = 2$, which coincides with the true DGP. As contamination increases, the minimum shifts leftward. Moreover, the loss surface becomes increasingly steep for values of γ_{fil} larger than the minimum, indicating heightened sensitivity of non-

robust updates to extreme realizations.

These findings can be interpreted through the lens of Theorem 1 and 2. Under MSE evaluation, a more robust update can be P_t–contractively dominating a less robust update when the update magnitude is sufficiently large. In the present volatility setting, large update magnitudes correspond to realizations of y_t^2 that are extreme relative to the current conditional variance $\vartheta_{t|t-1}$. The observational outliers in the training sample generate precisely such extreme squared-return realizations. While choosing $\gamma_{\text{fil}} < 2$ introduces a small bias relative to the quadratic target, it substantially reduces the variability of the estimated parameters under contamination. The resulting reduction in variance more than offsets the increase in squared bias, leading to a lower generalization error. By down-weighting extreme updates during estimation, the robust specification therefore yields volatility forecasts that remain closer to the latent variance path in the evaluation sample.

Figure 3: Robustness and forecast accuracy under a contaminated training sample



NOTE: Average out-of-sample mean squared error of the volatility forecasts as a function of γ_{fil}, for different contamination levels ε.

6 Conclusion

This paper introduces the BLADE filter, combining the PRADA class of De Punder (2026) with the conditional moment QSD filter of Blasques et al. (2023) by incorporating the adaptive loss of Barron (2019) into the parameter updating mechanism. The robustness of the update is governed by a continuous shape parameter γ which is learned from the data rather than fixed by a prespecified loss.

Under mild regularity conditions, the BLADE filter generates a strictly stationary, ergodic, and invertible sequence of time-varying parameters, and the associated quasi-likelihood estimator of the static parameters is consistent and asymptotically normal. We obtain these results by extending the QSD estimation framework of Blasques et al. (2023, Theorem 1) through a generalization of its required moment condition, which accommodates the Barron-loss-based estimating equation.

The BLADE filter belongs to the PRADA class, ensuring that an update reduces a local divergence from the true distribution in expectation. To compare update rules, we introduce P_t-contractive dominance, which orders two updates by their expected local divergence reduction towards P_t. Section 3.1 characterizes an explicit interval of learning rates over which a more robust BLADE update P_t-contractively dominates a less robust one under non-contaminated data, and we extend this to an explicit interval of contamination proportions under a contaminated updating draw, while both updates remain PRADA.

Monte Carlo experiments confirm the finite-sample relevance of these results. In non-contaminated samples the quadratic case ($\gamma = 2$) performs competitively, in line with strict consistency. Under contamination, the loss-minimizing γ departs systematically from 2, and values $\gamma < 2$ yield lower out-of-sample loss than both fixed-γ QSD filters and the benchmark GARCH and β_t–GARCH(1,1) models. This result is in line with Section 3.1, where more robust Barron updates achieve larger expected divergence reductions across a

wider range of step sizes.

References

Andersen, T. G. and T. Bollerslev (1998), "Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts", *International Economic Review*, *39*(4), 885.

Artemova, M., F. Blasques, J. van Brummelen, and S. J. Koopman (2022a). "Score-driven models: Methodology and theory". In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press, Oxford.

Artemova, M., F. Blasques, J. van Brummelen, and S. J. Koopman (2022b). "Score-driven models: Methods and applications". In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press, Oxford.

Barron, J. T. (2019). "A General and Adaptive Robust Loss Function". In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4326–4334. IEEE.

Barron, J. T. (2025). "A Power Transform".

Bernoulli, D. (1760), "Essai d'une Nouvelle Analyse de la Mortalite Causee par la Petite Verole, et des Avantages de l'Inoculation Pour la Prevenir", *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, 1–45.

Beutner, E., Y. Lin, and A. Lucas (2026), "Consistency, distributional convergence, and optimality of time-varying parameters in score-driven models", *Journal of Econometrics*, *255*, 106218.

Blasques, F. (2019). *Advanced Econometric Methods: A Guide to Estimation and Inference for Nonlinear Dynamic Models*. A. Publications.

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

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

Calvet, L. E., V. Czellar, and E. Ronchetti (2015), "Robust Filtering", *Journal of the American Statistical Association*, *110*(512), 1591–1606.

Catania, L., A. C. Harvey, and A. Luati (2025), "Robust CDF-Filtering of a Location Parameter", *Journal of Time Series Analysis*, *In Press*, Early Access: https://dx.doi.org:/10.1111/jtsa.70026.

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

Comaniciu, D. and P. Meer (2002), "Mean shift: A robust approach toward feature space analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619.

Creal, D., S. J. Koopman, and A. Lucas (2013), "Generalized autoregressive 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.

De Punder, R. (2026). "Proper and Robust Autoregressive Derivative Adaptive Models". Tinbergen Institute Discussion Paper TI 2026-022/III, Tinbergen Institute.

De Punder, R., T. Dimitriadis, and R.-J. Lange (2026). "Expected Kullback-Leibler-based characterizations of score-driven updates". Version Number: 3.

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*, *In Press*, Early Access: https://dx.doi.org/10.1080/01621459.2025.2576189.

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, B. Van Os, and D. Van Dijk (2025). "Gradient-based filtering under misspecification: Stability and error bounds". Tinbergen Institute Discussion Paper 25-006/III, Tinbergen Institute.

Durbin, J. (1960), "Estimation of Parameters in Time-Series Regression Models", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *22*(1), 139–153.

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.

Godambe, V. P. (1960), "An Optimum Property of Regular Maximum Likelihood Estimation", *The Annals of Mathematical Statistics*, *31*(4), 1208–1211.

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

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

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation", *Journal of the American Statistical Association*, *69*(346), 383–393.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons.

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*. Cambridge University Press.

Harvey, A. C. (2022), "Score-driven time series models", *Annual Review of Statistics and Its Application*, *9*(1), 321–342.

Harvey, A. C. and T. Chakravarty (2008). "Beta-t-(E)GARCH". Working paper, Faculty of Economics, Cambridge University, Cambridge.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (2 ed.). Springer Series in Statistics. Springer, New York.

Huber, P. J. (1981). *Robust statistics*. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York.

Lange, R.-J., B. Van Os, and D. Van Dijk (2026), "Implicit score-driven filters for time-varying parameter models", *Journal of Econometrics*, *255*, 106251.

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

Muler, N., D. Peña, and V. J. Yohai (2009), "Robust estimation for ARMA models", *The Annals of Statistics*, *37*(2), 816–840.

Muler, N. and V. J. Yohai (2008), "Robust estimates for GARCH models", *Journal of Statistical Planning and Inference*, *138*(10), 2918–2940.

Muler, N. and V. J. Yohai (2013), "Robust estimation for vector autoregressive models", *Computational Statistics & Data Analysis*, *65*, 68–79.

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.

Selten, R. (1998), "Axiomatic Characterization of the Quadratic Scoring Rule", *Experimental Economics*, *1*(1), 43–62.

Straumann, D. and T. Mikosch (2006), "Quasi-Maximum-Likelihood Estimation in Conditionally Heteroscedastic Time Series: A Stochastic Recurrence Equations Approach", *The Annals of Statistics*, *34*(5), 2449–2495.

Tobin, J. (1958), "Estimation of relationships for limited dependent variables", *Econometrica*, *26*(1), 24–36.

Tucker, H. G. (1959), "A Generalization of the Glivenko-Cantelli Theorem", *The Annals of Mathematical Statistics*, *30*(3), 828–830.

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 vol.3.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Appendix

A Proofs of the main results

A.1 Proof of Theorem 1

Proof. Let $(Y_t, \check{Y}_t) \sim \mathrm{P}_t^{\otimes 2}$ be independent copies. For $i \in \{1, 2\}$, denote the BLADE update, $\vartheta_{t|t}(\gamma_i, \check{Y}_t) \coloneqq \vartheta_{t|t-1} + \alpha\, \psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)$.

We first obtain an upper bound on the learning rate for the Barron updates to be PRADA (De Punder 2026, Theorem 1). For any $y, \check{y} \in \mathcal{Y}$, the path-integral version of the mean-value theorem applied to $S(F_{\vartheta_{t|t}(\gamma_i,\check{y}))}, y)$ at $\vartheta_{t|t-1}$ is given by

$$
\begin{aligned}
&S(\mathrm{F}_{\vartheta_{t|t}(\gamma_i,\check{y}))}, y) - S(\mathrm{F}_{\vartheta_{t|t-1}}, y) \\
&= (\vartheta_{t|t}(\gamma_i, \check{y}) - \vartheta_{t|t-1}) \nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, y) \\
&\qquad + (\vartheta_{t|t}(\gamma_i, \check{y}) - \vartheta_{t|t-1})^2 \int_0^1 (1-u) H_S(y, (1-u)\vartheta_{t|t-1} + u\vartheta_{t|t}(\gamma_i, \check{y}))\mathrm{d}u.
\end{aligned}
$$

Where $\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, y) \coloneqq \frac{\partial S(\mathrm{F}_\vartheta, y)}{\partial \vartheta}\Big|_{\vartheta=\vartheta_{t|t-1}}$ and $H_S(y, \vartheta) \coloneqq \frac{\partial^2 S(\mathrm{F}_\vartheta, y)}{\partial \vartheta^2}$. Using this expression in the expected divergence difference (1) and using (1) we obtain

$$
\begin{aligned}
\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i) &\coloneqq \mathbb{D}_S^{\mathrm{P}_t}(\mathrm{P}_t \| \mathrm{F}_{t|t-1}) - \mathbb{E}_{\check{Y}_t \sim \mathrm{P}_t} \mathbb{D}_S^{\mathrm{P}_t}(\mathrm{P}_t \| \mathrm{F}_{t|t}) \\
&= \mathbb{E}_{\mathrm{P}_t}\left[S(\mathrm{P}_t, Y_t^k) - S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\right] - \left(\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\left[S(\mathrm{P}_t, Y_t^k) - S(\mathrm{F}_{\vartheta_{t|t}(\gamma_i, \check{Y}_t^k)}, Y_t^k)\right]\right) \\
&= \mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\left[S(\mathrm{F}_{\vartheta_{t|t}(\gamma_i, \check{Y}_t^k)}, Y_t^k) - S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\right].
\end{aligned}
$$

Using the path-integral mean-value theorem expansion yields, the following representation

of the expected local S-divergence reduction

$$\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i) = \alpha\,\mathbb{E}_{\mathrm{P}_t}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big]\,\mathbb{E}_{\mathrm{P}_t}\Big[\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)\Big] + \alpha^2\,\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\,\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i))\Big], \tag{A.1}$$

where the integrated Hessian is exact:

$$\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i)) \coloneqq \int_0^1 (1-u)\,H_{\mathrm{S}}\Big(Y_t^k,\ \vartheta_{t|t-1} + u\alpha\,\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)\Big)\,\mathrm{d}u. \tag{A.2}$$

Here we used that Y_t and $\check{Y}_t$ are independent and that $\vartheta_{t|t}(\gamma_i, \check{y}_t) - \vartheta_{t|t-1} = \psi_k^{\mathrm{B}}(\check{y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)$. Similar to Gorgi et al. (2024, Assumption 3), we use that by Assumption 1(iii), there exists a $c \in (0, \infty)$ such that $-c \le \mathbb{E}_{\mathrm{P}_t}\left[\frac{\partial^2 S(\mathrm{F}_\vartheta, Y_t^k)}{\partial \vartheta^2}\right] \le 0$ for all $\vartheta \in \Theta$. This implies in particular that $\sup_{\vartheta\in\Theta}\left|\mathbb{E}_{\mathrm{P}_t}[H_{\mathrm{S}}(Y_t^k, \vartheta)]\right| \le c$, which is used to bound the path-integrated Hessian

$$\begin{aligned}
&\Big|\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi^2)\,\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i))\Big]\Big| \\
&= \left|\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\left[\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2 \int_0^1 (1-u)\,H_{\mathrm{S}}\Big(Y_t^k,\ \vartheta_{t|t-1} + u\alpha\,\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)\Big)\,\mathrm{d}u\right]\right| \\
&\le \mathbb{E}_{\mathrm{P}_t}\left[\left|\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\right| \left|\int_0^1 (1-u)\,\mathbb{E}_{\mathrm{P}_t}\Big[H_{\mathrm{S}}\Big(Y_t^k,\ \vartheta_{t|t-1} + u\alpha\,\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)\Big)\Big]\,\mathrm{d}u\right|\right] \\
&\le \mathbb{E}_{\mathrm{P}_t}\left[\left|\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\right| \int_0^1 (1-u)\,\left|\mathbb{E}_{\mathrm{P}_t}\Big[H_{\mathrm{S}}\Big(Y_t^k,\ \vartheta_{t|t-1} + u\alpha\,\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)\Big)\Big]\right|\,\mathrm{d}u\right] \\
&\le \mathbb{E}_{\mathrm{P}_t}\left[\left|\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\right| \int_0^1 (1-u)\,\sup_{\vartheta\in\Theta}\left|\mathbb{E}_{\mathrm{P}_t}\Big[H_{\mathrm{S}}\Big(Y_t^k,\ \vartheta\Big)\Big]\right|\,\mathrm{d}u\right] \\
&= \frac{c}{2}\mathbb{E}_{\mathrm{P}_t}\Big[\Big|\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\Big|\Big] \\
&= \frac{c}{2}\mathbb{E}_{\mathrm{P}_t}\Big[\psi_k^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\Big].
\end{aligned}$$

Where the third line results from independence of Y_t^k and $\check{Y}_t^k$ under $\mathrm{P}_t^{\otimes 2}$ and by using that for any random variable Z, $|\mathbb{E}[Z]| \le \mathbb{E}[|Z|]$. Assumption 1(iv) ensures finiteness. This

allows us to conclude that

$$\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i) \geq \alpha\, \mathbb{E}_{\mathrm{P}_t}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big]\, \mathbb{E}_{\mathrm{P}_t}\big[\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)\big] - \frac{c}{2}\,\alpha^2\, \mathbb{E}_{\mathrm{P}_t}\Big[\psi_k^{\mathrm{B}}(Y_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\Big],$$

which is strictly positive whenever

$$0 < \alpha < \frac{2}{c}\frac{\mathbb{E}_{\mathrm{P}_t}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big]\, \mathbb{E}_{\mathrm{P}_t}\big[\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)\big]}{\mathbb{E}_{\mathrm{P}_t}\big[\psi_k^{\mathrm{B}}(Y_t^k, \vartheta_{t|t-1}; \gamma_i, \xi)^2\big]} =: \bar{\alpha}_t(\gamma_i),$$

where we use gradient alignment to conclude that the upper bound is positive (Assumption (1)(v)). Taking the minimum over $i \in \{1, 2\}$ gives $\bar{\alpha}_t = \min_i \bar{\alpha}_t(\gamma_i)$, so $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i) > 0$ for both $i \in \{1, 2\}$ and all $\alpha \in (0, \bar{\alpha}_t)$.

Now we obtain a lower bound on α for the robust update to induce a larger divergence reduction. Define

$$g(\vartheta_1, \vartheta_2) \coloneqq S(\mathrm{F}_{\vartheta_1}, Y_t^k) - S(\mathrm{F}_{\vartheta_2}, Y_t^k),$$

so that $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) = \mathbb{E}_{\mathrm{P}_t^{\otimes 2}}[g\big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_1), \vartheta_{t|t}(\check{Y}_t^k, \gamma_2)\big)]$. Applying a second-order Taylor expansion of g in the vector $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)$ around $\boldsymbol{\vartheta}_0 = (\vartheta_{t|t-1}, \vartheta_{t|t-1})$ gives

$$\begin{aligned} g\big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_1), \vartheta_{t|t}(\check{Y}_t^k, \gamma_2)\big) = {} & \nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big[\big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_1) - \vartheta_{t|t-1}\big) - \big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_2) - \vartheta_{t|t-1}\big)\Big] \\ & + \big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_1) - \vartheta_{t|t-1}\big)\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_1) - \vartheta_{t|t-1}\big) \\ & - \big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_2) - \vartheta_{t|t-1}\big)\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\big(\vartheta_{t|t}(\check{Y}_t^k, \gamma_2) - \vartheta_{t|t-1}\big). \end{aligned} \tag{A.3}$$

Substituting $\vartheta_{t|t}(\check{Y}_t^k, \gamma_i) - \vartheta_{t|t-1} = \alpha\, \psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)$ into (A.3) and taking expectations

under $\mathrm{P}_t^{\otimes 2}$ yields,

$$\begin{aligned}\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) &= \alpha\,\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big(\psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}) - \psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big)\Big] \\ &\quad + \alpha^2\,\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\,\psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2 \qquad \text{(A.4)} \\ &\quad - \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\,\psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\Big] \\ &= -\alpha\,A_t + \alpha^2\,Q_t, \qquad \text{(A.5)}\end{aligned}$$

where $\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))$ is given in (A.2). With

$$A_t \coloneqq \mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\big(\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_2, \xi) - \psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_1, \xi)\big)\Big], \qquad \text{(A.6)}$$

and

$$\begin{aligned}Q_t \coloneqq &\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\,\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_1, \xi)^2 \\ &- \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\,\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_2, \xi)^2\Big]. \qquad \text{(A.7)}\end{aligned}$$

We now show $Q_t \geq 0$. For any $\gamma \leq 2$, the Barron score admits the representation

$$\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma, \xi) = W_t'(\gamma)\,\psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma{=}2, \xi),$$

where

$$W_t'(\gamma) \coloneqq \left(1 + \frac{(\check{Y}_t^k - \vartheta_{t|t-1})^2}{\xi^2|\gamma - 2|}\right)^{\gamma/2-1}.$$

Since $\gamma/2 - 1$ is increasing in γ and $1 + (\cdot)^2 \geq 1$, a smaller values of γ gives a smaller exponent, so $\gamma_1 \leq \gamma_2$ implies $W_t'(\gamma_1) \leq W_t'(\gamma_2)$ a.s., with both weights in $(0, 1]$ a.s.

Define, for each fixed $\check{Y}_t^k$ and $i \in \{1,2\}$,

$$\bar{H}_i(u) \coloneqq \mathbb{E}_{\mathrm{P}_t}\Big[H_{\mathrm{S}}\Big(Y_t^k,\, \vartheta_{t|t-1} + u\alpha\, \psi_k^{\mathrm{B}}(\check{Y}_t, \vartheta_{t|t-1}; \gamma_i, \xi)\Big)\Big],$$

which is finite for all $u \in [0,1]$ a.s. and satisfies $\bar{H}_i(u) \leq 0$ for all $u \in [0,1]$ by Assumptions 1(ii),(iii), since $\vartheta_{t|t-1} + u\alpha\psi_k^{\mathrm{B}}(\gamma_i) \in \Theta$ a.s. By Assumption 1(iv), Fubini's theorem gives

$$\mathbb{E}_{\mathrm{P}_t}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i))\Big] = \int_0^1 (1-u)\, \bar{H}_i(u)\, \mathrm{d}u.$$

Using $\psi_k^{\mathrm{B}}(\gamma_i) = W_t'(\gamma_i)\psi_k^{\mathrm{B}}(\gamma{=}2)$ and performing the substitution $s = uW_t'(\gamma_i)$ for each i,

$$(W_t'(\gamma_i))^2 \mathbb{E}_{\mathrm{P}_t}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i))\Big] = \int_0^{W_t'(\gamma_i)} (W_t'(\gamma_i) - s)\, \bar{H}^*(s)\, \mathrm{d}s, \tag{A.8}$$

where $\bar{H}^*(s) \coloneqq \mathbb{E}_{\mathrm{P}_t}[H_{\mathrm{S}}(Y_t^k, \vartheta_{t|t-1} + s\alpha\psi_k^{\mathrm{B}}(\gamma{=}2))] \leq 0$ is the same function for both i, since both paths are parametrized through the common reference score $\psi_k^{\mathrm{B}}(\gamma{=}2)$.

We show $(W_t'(\gamma_1))^2 \mathbb{E}_{\mathrm{P}_t}[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))] \geq (W_t'(\gamma_2))^2 \mathbb{E}_{\mathrm{P}_t}[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))]$ via two inequalities. Since $\bar{H}^*(s) \leq 0$ and $0 \leq W_t'(\gamma_1) - s \leq W_t'(\gamma_2) - s$ for $s \in [0, W_t'(\gamma_1)]$ a.s., replacing the weight $W_t'(\gamma_1) - s$ by the larger $W_t'(\gamma_2) - s$ makes the integral more negative,

$$\int_0^{W_t'(\gamma_1)} (W_t'(\gamma_1) - s)\, \bar{H}^*(s)\, \mathrm{d}s \geq \int_0^{W_t'(\gamma_1)} (W_t'(\gamma_2) - s)\, \bar{H}^*(s)\, \mathrm{d}s. \tag{A.9}$$

Since $\bar{H}^*(s) \leq 0$, extending the integration domain from $[0, W_t'(\gamma_1)]$ to $[0, W_t'(\gamma_2)]$ only adds non-positive contributions, so

$$\begin{aligned} &\int_0^{W_t'(\gamma_1)} (W_t'(\gamma_2) - s)\, \bar{H}^*(s)\, \mathrm{d}s \\ &\geq \int_0^{W_t'(\gamma_2)} (W_t'(\gamma_2) - s)\, \bar{H}^*(s)\, \mathrm{d}s = (W_t'(\gamma_2))^2\, \mathbb{E}_{\mathrm{P}_t}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\Big]. \end{aligned} \tag{A.10}$$

Combining (A.8), (A.9), and (A.10), one obtains

$$(W_t'(\gamma_1))^2\,\mathbb{E}_{\mathrm{P}_t}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k,\vartheta_{t|t}(\check{Y}_t^k,\gamma_1))\Big] \geq (W_t'(\gamma_2))^2\,\mathbb{E}_{\mathrm{P}_t}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k,\vartheta_{t|t}(\check{Y}_t^k,\gamma_2))\Big].$$

Multiplying both sides by $\psi_k^{\mathrm{B}}(\check{Y}_t,\vartheta_{t|t-1};\gamma{=}2,\xi)^2 \geq 0$ and using $\psi_k^{\mathrm{B}}(\gamma_i) = W_t'(\gamma_i)\psi_k^{\mathrm{B}}(\gamma{=}2)$ followed by taking expectations over $\check{Y}_t^k$ under P_t,

$$\mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k,\vartheta_{t|t}(\check{Y}_t^k,\gamma_1))\,\psi_1^{\mathrm{B}}(\check{Y}_t^k,\vartheta_{t|t-1})^2 - \widetilde{H}_{\mathrm{S}}(Y_t^k,\vartheta_{t|t}(\check{Y}_t^k,\gamma_2))\,\psi_2^{\mathrm{B}}(\check{Y}_t^k,\vartheta_{t|t-1})^2\Big] \geq 0,$$

hence $Q_t \geq 0$.

Since $Q_t \geq 0$, the right-hand side of (A.5) is a convex parabola in $\alpha \geq 0$. If $A_t \leq 0$, every term is non-negative and $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \geq 0$ for all $\alpha \geq 0$. If $A_t > 0$, then $\psi_k^{\mathrm{B}}(\check{Y}_t,\vartheta_{t|t-1};\gamma{=}2) \neq 0$ under P_t (otherwise $A_t = 0$) , and Assumption 1(iii) yields $Q_t > 0$. The parabola then crosses zero at $\underline{\alpha}_t := A_t/Q_t > 0$, and $-\alpha A_t + \alpha^2 Q_t \geq 0$ for all $\alpha \geq \underline{\alpha}_t$.

Combining with the PRADA upper bound $\bar{\alpha}_t$, both $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i) > 0$ and $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \geq 0$ hold simultaneously for all $\alpha \in [\underline{\alpha}_t, \bar{\alpha}_t)$, this completes the proof. □

A.2 Proof of Theorem 2

Proof. From Definition 1 we obtain the following expression for the divergence reduction of a BLADE update ϕ_i^{B}, $\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}}) = \mathbb{D}_S(\mathrm{P}_t\|\mathrm{F}_{t|t-1}) - \mathbb{E}_{\check{Y}_t\sim\mathrm{P}_t^\varepsilon}\mathbb{D}_S(\mathrm{P}_t\|\mathrm{F}_{t|t})$. Following the proof of Theorem 1 expand $\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}})$

$$\begin{aligned}\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}}) = \alpha\,\mathbb{E}_{\mathrm{P}_t}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}},Y_t^k)\Big]\,\mathbb{E}_{\mathrm{P}_t^\varepsilon}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k,\vartheta_{t|t-1})\Big] \\ + \alpha^2\,\mathbb{E}_{\mathrm{P}_t\otimes\mathrm{P}_t^\varepsilon}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k,\vartheta_{t|t-1})^2\,\widetilde{H}_{\mathrm{S}}(Y_t^k,\vartheta_{t|t}(\check{Y}_t^k,\gamma_i))\Big].\end{aligned}$$

By independence of $Y_t^k \sim \mathrm{P}_t$ and $\check{Y}_t^k \sim \mathrm{P}_t^\varepsilon$ and using linearity of the expectation, expanding $\mathrm{P}_t^\varepsilon = (1-\varepsilon)\mathrm{P}_t + \varepsilon \mathrm{H}_t$ yields

$$\begin{aligned}
\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}}) = \alpha\, \mathbb{E}_{\mathrm{P}_t}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big] \Big((1-\varepsilon)\, \mathbb{E}_{\mathrm{P}_t}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big] + \varepsilon\, \mathbb{E}_{\mathrm{H}_t}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big]\Big) \\
+ \alpha^2 \Bigg((1-\varepsilon)\, \mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\, \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i))\Big] \\
+ \varepsilon\, \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{H}_t}\Big[\psi_i^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\, \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_i))\Big] \Bigg). \\
= (1-\varepsilon)\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) + \varepsilon\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}})
\end{aligned}$$

For both updates to remain PRADA we must have that $\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}}) > 0$. Under Assumption 1 both updates are PRADA under P_t, so $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) > 0$ for $i \in \{1,2\}$. Using the mixture decomposition $\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}}) = (1-\varepsilon)\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) + \varepsilon\, \Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}})$, two cases arise. If $\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}}) \geq 0$, the convex combination is strictly positive for all $\varepsilon \in [0,1]$ and the PRADA condition is satisfied regardless of the contamination level. If $\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}}) < 0$, requiring $\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}}) > 0$ yields

$$\varepsilon \;<\; \frac{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}})}{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) - \Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}})},$$

where the right-hand side lies in $(0,1)$. Combining the two cases results in,

$$\bar{\varepsilon}(\gamma_i, \alpha) = \min\left\{ \frac{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}})}{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) - \Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}})}, 1 \right\}.$$

Requiring both updates to remain PRADA simultaneously gives the overall upper bound

$$\bar{\varepsilon} \;\coloneqq\; \min_{i \in \{1,2\}}\; \bar{\varepsilon}(\gamma_i, \alpha).$$

We now establish a lower bound on the contamination level ε for the robust update to be P_t–contractively dominating the less robust update. Recall from Definition 3 that the

expected divergence reduction difference of two BLADE updates, $\phi_1^{\mathrm{B}}, \phi_2^{\mathrm{B}}$, which are updates using the contaminated measure $\mathrm{P}_t^{\varepsilon} = (1-\varepsilon)\mathrm{P}_t + \varepsilon\,\mathrm{H}_t$ (8), is given by

$$\begin{aligned}
\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) &= \alpha\,\mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t^{\varepsilon}}\Big[\nabla_{\vartheta} S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big(\psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}) - \psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big)\Big] \\
&\quad + \alpha^2\,\mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t^{\varepsilon}}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\,\psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2 \\
&\quad - \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\,\psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\Big] \\
&= -\alpha\,A_t + \alpha^2\,Q_t.
\end{aligned}$$

Importantly, we apply this contamination only to the updating draw, so the two integrating measures are P_t (for Y_t^k) and $\mathrm{P}_t^{\varepsilon}$ (for $\check{Y}_t^k$). Under this product measure, the expansion can be written as,

$$\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) = -\alpha\,A_t^{\varepsilon} + \alpha^2\,Q_t^{\varepsilon}, \tag{A.11}$$

where

$$\begin{aligned}
A_t^{\varepsilon} &\coloneqq \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t^{\varepsilon}}\Big[\nabla_{\vartheta} S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big(\psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}) - \psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big)\Big], \\
Q_t^{\varepsilon} &\coloneqq \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t^{\varepsilon}}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\,\psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2 - \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\,\psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\Big].
\end{aligned}$$

Since $Y_t^k \sim \mathrm{P}_t$ and $\check{Y}_t^k \sim \mathrm{P}_t^{\varepsilon}$ are independent and enter different factors of the integrands, expanding $\mathrm{P}_t^{\varepsilon} = (1-\varepsilon)\mathrm{P}_t + \varepsilon\,\mathrm{H}_t$ yields

$$\begin{aligned}
A_t^{\varepsilon} &= (1-\varepsilon)\,A_t + \varepsilon\,A_t^H, \\
Q_t^{\varepsilon} &= (1-\varepsilon)\,Q_t + \varepsilon\,Q_t^H,
\end{aligned}$$

where A_t, Q_t are the non-contaminated quantities computed under $\mathrm{P}_t^{\otimes 2}$, and

$$A_t^H \coloneqq \mathbb{E}_{\mathrm{P}_t}\left[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\right]\, \mathbb{E}_{\mathrm{H}_t}\left[\psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}) - \psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\right],$$

$$Q_t^H \coloneqq \mathbb{E}_{\mathrm{P}_t\otimes\mathrm{H}_t}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\, \psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2 - \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\, \psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\Big].$$

Using these expressions we obtain the following expression for $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}})$,

$$\begin{aligned}\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) &= -\alpha\big((1-\varepsilon)\, A_t + \varepsilon\, A_t^H\big) + \alpha^2\big((1-\varepsilon)\, Q_t + \varepsilon\, Q_t^H\big)\\ &= -\alpha(1-\varepsilon)\, A_t - \alpha\varepsilon\, A_t^H + \alpha^2(1-\varepsilon)\, Q_t + \alpha^2\varepsilon\, Q_t^H\\ &= (1-\varepsilon)\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) + \varepsilon\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}). \qquad \text{(A.12)}\end{aligned}$$

Assuming that $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t} \geq 0$, we separate two cases on $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}})$. If $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \geq 0$ the inequality in (A.12) is greater than 0 for all $\varepsilon \in [0,1]$. In the other case that $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \leq 0$, solving (A.12) for ε, we obtain the following bound on ε for which $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \geq 0$,

$$\varepsilon \geq \underline{\varepsilon} \coloneqq \frac{-\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}})}{\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) - \Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}})},$$

where $\varepsilon \in [0,1]$. Combining the two cases give the lower bound,

$$\underline{\varepsilon} \coloneqq \max\left\{0, \frac{-\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}})}{\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) - \Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}})}\right\},$$

which for all $\varepsilon \geq \underline{\varepsilon}$ result in $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_1^{\mathrm{B}}\|\phi_2^{\mathrm{B}}) \geq 0$.

Combining the upper and lower bound on ε, we conclude that for $\varepsilon \in (\underline{\varepsilon}, \bar{\varepsilon})$ the robust update induces a larger divergence reduction under contamination and both updates remain PRADA. □

A.3 Proofs for consistency and asymptotic normality of the QLE estimates

We state Lemmas 1–4 of Blasques et al. (2023) needed for Theorem 3 and 4; they are verified in Appendix D. Following the notation of Blasques et al. (2023), denote the exogenous variables, $z_t = (\varepsilon_t, X_t')' \in \mathbb{R}^d$ and note that the time-varying parameter $\vartheta_{t|t-1}$ satisfies a recurrence relation of the form $\vartheta_{t+1|t} = \varphi(z_t, \vartheta_{t|t-1})$. Conditions for the QSD model to generate a stationary sequence have been proposed by Blasques et al. (2023) in their Lemma 1.

Lemma A.1. *Assume that (z_t) is stationary and ergodic. Suppose that*

$$\textit{(i)} \quad \mathbb{E}_{\bar{\mathrm{P}}}\Big[\log^+\big|\psi^{\mathrm{B}}_{\gamma,\xi}\big(g(\vartheta^0,\varepsilon_t),\vartheta^0\big)\big|\Big] < \infty, \quad \vartheta^0 \in \Theta \subset \mathbb{R}; \tag{A.13}$$

$$\textit{(ii)} \quad \mathbb{E}_{\bar{\mathrm{P}}}\big[\log \mathfrak{L}_t\big] < 0, \quad \mathfrak{L}_t = \sup_{\vartheta\in\Theta}\Big|\gamma\,\frac{\partial}{\partial\vartheta}\psi^{\mathrm{B}}_{\gamma,\xi}\big(g(\vartheta,\varepsilon_t),\vartheta\big) + \beta\Big|, \tag{A.14}$$

Then there exist unique strictly stationary and ergodic solutions to $\{\vartheta_{t|t-1}\}_{t\in\mathbb{Z}}$ and $\{y_t\}_{t\in\mathbb{Z}}$ to $y_t = g(\vartheta_{t|t-1},\varepsilon_t)$ and $\vartheta_{t|t-1}$ follows (4).

A sufficient condition for (i) is $\mathbb{E}_{\bar{\mathrm{P}}}\log^+|Y_t| < \infty$, since $|\psi^{\mathrm{B}}_{\gamma,\xi}|$ is uniformly bounded for $\gamma \le 1$ and bounded by $|x_k|/\xi^2$ for $1 < \gamma \le 2$. For condition (ii) a sufficient condition is that

$$\frac{|\alpha|}{\xi^2} + |\beta| < 1.$$

Lemma 2 in Blasques et al. (2023) gives conditions for unconditional bounded moments.

Lemma A.2. *Under the assumptions of Lemma A.1 hold, if the sequence $\{\mathfrak{L}_t\}$ is i.i.d. and for some $r > 0$,*

$$\textit{(i)}\ \mathbb{E}_{\bar{\mathrm{P}}}\left[\big|\psi^{\mathrm{B}}_k\big(g(\vartheta^0,\varepsilon_t),\vartheta^0;\gamma,\xi\big)\big|^r\right] < \infty \quad \textit{and} \quad \textit{(ii)}\ \mathbb{E}_{\bar{\mathrm{P}}}\left[\sup_{\vartheta\in\Theta}\Big|\frac{\partial\psi^{\mathrm{B}}_k\big(g(\vartheta,\varepsilon_t),\vartheta;\gamma,\xi\big)}{\partial\vartheta}\Big|^r\right] < \infty,$$

for some $\vartheta^0 \in \Theta$. *Then* $\vartheta_{t|t-1}$ *satisfies* $\mathbb{E}_{\bar{\mathrm{P}}} |\vartheta_{t|t-1}|^s < \infty$ *for some* $s > 0$.

These conditions hold in the proposed model if $\gamma \leq 2$ and if $\mathbb{E}_{\bar{\mathrm{P}}}|x_k(y_t, \vartheta_{t|t-1})|$ is bounded.

To show that the BLADE filter is invertible, the following condition from Blasques et al. (2023) is shown to hold,

Lemma A.3. *Let* $\{y_t, X_t\}_{t\in\mathbb{Z}}$ *be stationary and ergodic, and suppose that*

(i) for all $\lambda \in \Lambda$, *there exists* $\vartheta^0 \in \Theta$ *such that* $\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\left|\psi^{\mathrm{B}}_k(y_t, \vartheta^0; \gamma, \xi)\right|\right] < \infty$,

(ii) $\mathbb{E}_{\bar{\mathrm{P}}}\Big[\log \sup_{\vartheta\in\Theta} \sup_{\lambda\in\Lambda}\Big|\alpha \frac{\partial}{\partial\vartheta}\psi^{\mathrm{B}}_{\gamma,\xi}(y_t, \vartheta) + \beta\Big|\Big] < 0.$

Then for all $\lambda \in \Lambda$, *there exists a unique stationary and ergodic solution to* $\{\vartheta_{t|t-1}(\lambda)\}_{\lambda\in\Lambda}$ *to* $\vartheta_{t+1|t}(\lambda) = \omega + \alpha\psi^{\mathrm{B}}_{\gamma,\xi}(y_t, \vartheta_{t|t-1}(\lambda)) + \beta\vartheta_{t|t-1}(\lambda),\ t \in \mathbb{Z}$. *Furthermore, for all starting functions* $\hat{\vartheta}_1(\cdot) \in \mathbb{C}(\Lambda, \Theta)$, *there exists* $\varrho \in (0, 1)$ *such that*

$$\varrho^{-t} \sup_{\lambda\in\Lambda}\left|\hat{\vartheta}_{t|t-1}(\lambda) - \vartheta_{t|t-1}(\lambda)\right| \to 0 \quad a.s.\ as \quad (t \to \infty).$$

The model is then said to be uniformly invertible.

Under similar assumptions as in Lemma A.1 the two conditions of Lemma A.3 hold. The sufficient assumptions are, $\gamma \leq 2$ and $\mathbb{E}_{\bar{\mathrm{P}}} \log^+ |Y_t| < \infty$.

This proposition sets the conditions for stationarity and invertibility of the derivatives of the filter. Define,

$$\vartheta''_{t+1|1}(\lambda) \coloneqq \mathrm{vec}\left(\frac{\partial^2 \vartheta_{t+1|t}}{\partial\lambda\,\partial\lambda^\top}\right) = C_t + b_t\,\vartheta''_{t|t-1}(\lambda), \tag{A.15}$$

where

$$
\begin{aligned}
b_t &= \alpha\,\frac{\partial \psi_k^{\mathrm{B}}}{\partial \vartheta} + \beta, \\
C_t &= \operatorname{vec}\Bigg(\frac{\partial \alpha}{\partial \lambda}\frac{\partial \psi_k^{\mathrm{B}}}{\partial \lambda^\top} + \frac{\partial \psi_k^{\mathrm{B}}}{\partial \vartheta}\frac{\partial \alpha}{\partial \lambda}\big(\vartheta_t'\big)^\top + \frac{\partial \psi_k^{\mathrm{B}}}{\partial \lambda}\frac{\partial \alpha}{\partial \lambda^\top} + \alpha\frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \vartheta\,\partial \lambda}\big(\vartheta_t'\big)^\top + \alpha\frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \lambda\,\partial \lambda^\top} \\
&\qquad + \frac{\partial \beta}{\partial \lambda}\big(\vartheta_t'\big)^\top + \frac{\partial \psi_k^{\mathrm{B}}}{\partial \vartheta}\,\vartheta_t'\frac{\partial \alpha}{\partial \lambda^\top} + \alpha\,\vartheta_t'\frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \vartheta\,\partial \lambda^\top} + \alpha\frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \vartheta^2}\vartheta_t'\big(\vartheta_t'\big)^\top \Bigg),
\end{aligned}
$$

with $\vartheta_t' := \frac{\partial \vartheta_{t|t-1}}{\partial \lambda}$ and $\psi_k^{\mathrm{B}} = \psi_k^{\mathrm{B}}(y_t, \vartheta_{t|t-1}; \gamma, \xi)$. Using $x_k(y, \vartheta) = y^k - \vartheta$ so that $\frac{\partial x_k}{\partial \vartheta} = -1$ and $\frac{\partial^2 x_k}{\partial \vartheta^2} = 0$, the score derivatives simplify to

$$
\frac{\partial \psi_k^{\mathrm{B}}}{\partial \vartheta} = -\frac{\partial^2 \tilde{S}_{\gamma,\xi}^{\mathrm{B}}}{\partial x_k^2}, \qquad \frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \vartheta^2} = \frac{\partial^3 \tilde{S}_{\gamma,\xi}^{\mathrm{B}}}{\partial x_k^3}, \qquad \frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \vartheta\,\partial \lambda} = -\frac{\partial^3 \tilde{S}_{\gamma,\xi}^{\mathrm{B}}}{\partial x_k^2\,\partial \lambda},
$$

where all derivatives of $\tilde{S}_{\gamma,\xi}^{\mathrm{B}}$ are evaluated at $x_k(y_t, \vartheta_{t|t-1})$. The derivative of $\hat{\vartheta}_{t|t-1}(\lambda)$ w.r.t. λ is given by

$$
\frac{\partial \hat{\vartheta}_{t+1|t}(\lambda)}{\partial \lambda} = \frac{\partial \omega}{\partial \lambda} + \psi_{\gamma,\xi}^{\mathrm{B}}\frac{\partial \alpha}{\partial \lambda} + \alpha\frac{\partial \psi_{\gamma,\xi}^{\mathrm{B}}}{\partial \lambda} + \hat{\vartheta}_{t|t-1}(\lambda)\frac{\partial \beta}{\partial \lambda} + \left(\alpha\frac{\partial \psi_{\gamma,\xi}^{\mathrm{B}}}{\partial \vartheta} + \beta\right)\frac{\partial \hat{\vartheta}_{t|t-1}(\lambda)}{\partial \lambda}, \quad \text{(A.16)}
$$

where $\psi_{\gamma,\xi}^{\mathrm{B}} \equiv \psi_k^{\mathrm{B}}(y_t, \hat{\vartheta}_{t|t-1}; \gamma, \xi)$.

Lemma A.4. *Let the conditions of Lemma A.3 hold, assume that $\psi_{\gamma,\xi}^{\mathrm{B}}$ admits continuous second-order derivatives with respect to its last two components, and suppose that*

$$
\textit{(i) for all } \lambda \in \Lambda, \quad \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+|\psi_k^{\mathrm{B}}| + \log^+\left\|\frac{\partial \psi_k^{\mathrm{B}}}{\partial \lambda}\right\| + \log^+\left|\frac{\partial \psi_k^{\mathrm{B}}}{\partial \vartheta}\right| + \log^+|\vartheta_{t|t-1}(\lambda)|\right] < \infty.
$$

Then, for all $\lambda \in \Lambda$, there exists a unique strictly stationary and ergodic solution $\{\vartheta_{t|t-1}'(\lambda)\}_{t\in\mathbb{Z}}$ to (A.16)*. If in addition*

$$
\textit{(ii)} \quad \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\left(\sup_{\vartheta\in\Theta}\left|\frac{\partial \psi_k^{\mathrm{B}}}{\partial \vartheta}\right| + \sup_{\vartheta,\lambda}\left\|\frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \lambda \partial \vartheta}\right\| + \sup_{\vartheta\in\Theta}\left\|\frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \vartheta^2}\right\| + \sup_{\lambda}\left\|\vartheta_{t|t-1}'(\lambda)\right\|\right)\right] < \infty,
$$

then, for all starting functions $\hat{\vartheta}_1(\cdot) \in \mathcal{C}(\Lambda, \Theta)$ *and* $\hat{\vartheta}'_1(\cdot) \in \mathcal{C}(\Lambda, \mathbb{R}^q)$*, there exists* $\varrho \in (0,1)$ *such that*

$$\varrho^{-t} \sup_{\lambda \in \Lambda} \left\| \hat{\vartheta}'_{t|t-1}(\lambda) - \vartheta'_{t|t-1}(\lambda) \right\| \to 0 \quad a.s.\ as\ t \to \infty.$$

Furthermore, assuming that

$$\textit{(iii) for all } \lambda \in \Lambda, \quad \mathbb{E}_{\bar{\mathrm{P}}} \left[\log^+ \left\| \frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \lambda \partial \lambda^\top} \right\| + \log^+ \left\| \frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \lambda \partial \vartheta} \right\| + \log^+ \left\| \frac{\partial^2 \psi_k^{\mathrm{B}}}{\partial \vartheta^2} \right\| \right] < \infty,$$

then, for all $\lambda \in \Lambda$*, there exists a unique strictly stationary and ergodic solution* $\{\vartheta''_{t|t-1}(\lambda)\}_{t \in \mathbb{Z}}$ *to* (A.15)*. Under the additional assumption*

$$\textit{(iv)} \quad \mathbb{E}_{\bar{\mathrm{P}}} \left[\log^+ \left(\sup_{\vartheta, \lambda_i, \lambda_j} \left| \frac{\partial^3 \psi_k^{\mathrm{B}}}{\partial \lambda_i \partial \lambda_j \partial \vartheta} \right| + \sup_{\vartheta, \lambda_i} \left\| \frac{\partial^3 \psi_k^{\mathrm{B}}}{\partial \lambda_i \partial \vartheta^2} \right\| + \sup_{\vartheta \in \Theta} \left\| \frac{\partial^3 \psi_k^{\mathrm{B}}}{\partial \vartheta^3} \right\| \right) \right] < \infty,$$

with $\psi_k^{\mathrm{B}} = \psi_k^{\mathrm{B}}(y_t, \vartheta_{t|t-1})$*, then, for all starting functions* $\hat{\vartheta}_1(\cdot) \in \mathcal{C}(\Lambda, \Theta)$*,* $\hat{\vartheta}'_1(\cdot) \in \mathcal{C}(\Lambda, \mathbb{R}^q)$*, and* $\hat{\vartheta}''_1(\cdot) \in \mathcal{C}(\Lambda, \mathbb{R}^{q^2})$*, there exists* $\varrho \in (0,1)$ *such that*

$$\varrho^{-t} \sup_{\lambda \in \Lambda} \left\| \hat{\vartheta}''_{t|t-1}(\lambda) - \vartheta''_{t|t-1}(\lambda) \right\| \to 0 \quad a.s.\ as \quad t \to \infty.$$

These conditions hold under mild assumptions. These include bounded derivatives of $x_k(y_t, \vartheta_{t|t-1})$ with respect to ϑ, up to and including the fourth derivative and $\gamma \in \Lambda \subseteq (-\infty, 2]$, $\xi > 0$ so that all partial derivatives of $S^{\mathrm{B}}_{\gamma,\xi}(x_k(y_t, \vartheta_{t|t-1}), \gamma, \xi)$ with respect to $x_k(y_t, \vartheta_{t|t-1})$ are bounded.

A.4 Proof of Theorem 3

Proof. This proof follows the same structure as the proof of Theorem 1 in Blasques et al. (2023). Under Assumption 3(i),(ii), $(Y_t, \mathcal{F}_t^{Y,Z})_{t \in \mathbb{Z}}$ is a strictly stationary and ergodic process and let $\Lambda \subseteq \mathbb{R}^p$ be compact. Recall that $\tilde{\mathrm{P}}_t$ denotes the true conditional distribution of Y_t given $\mathcal{F}_{t-1}^{Y,Z}$, and that the target functional ϑ_t solves $\mathbb{E}_{\tilde{\mathrm{P}}_t}[\psi^{\mathrm{B}}_{\gamma,\xi}(Y_t^k - \vartheta_t)] = 0$. Under Assump-

tion 3, for each $\lambda \in \Lambda$ the recursion (4) admits a unique stationary solution $\vartheta_{t|t-1}(\lambda) \in \mathbb{R}$. We maintain throughout that the filter is correctly specified for this functional: there exists $\lambda_0 \equiv \lambda_0(\gamma) \in \Lambda$ such that $\vartheta_{t|t-1}(\lambda_0) = \vartheta_t$ a.s. for all $t \in \mathbb{Z}$, so that the estimating function is conditionally unbiased, $\mathbb{E}_{\tilde{\mathrm{P}}_t}[h_t(\lambda_0)] = 0$. Let

$$\frac{\partial \vartheta_{t|t-1}(\lambda)}{\partial \lambda} \coloneqq \frac{\partial}{\partial \lambda^\top} \vartheta_{t|t-1}(\lambda) \in \mathbb{R}^{1\times p}.$$

Let $\psi^{\mathrm{B}}_{\gamma,\xi} : \mathbb{R} \to \mathbb{R}$ be the (Barron) function applied to the scalar residual $x_k = Y_t^k - \vartheta_{t|t-1}(\lambda)$ and $h_t = \psi^{\mathrm{B}}_{\gamma,\xi}\big(Y_t^k - \vartheta_{t|t-1}(\lambda)\big)$.

For each $\gamma \leq 2$, let $\lambda_0 \equiv \lambda_0(\gamma)$ be the pseudo truth, the moment condition under the true measure P_t holds,

$$\mathbb{E}_{\tilde{\mathrm{P}}_t}[h_t(\lambda_0)] = 0. \tag{A.17}$$

Define the (conditional) scaling

$$\sigma_t^2(\lambda) \coloneqq \mathbb{E}_{\tilde{\mathrm{P}}_t}\big[h_t(\lambda)^2\big],$$

then under Assumption 3(iii) there exists $\underline{\sigma}^2 > 0$ such that

$$\inf_{\lambda\in\Lambda} \sigma_t^2(\lambda) \geq \underline{\sigma}^2. \tag{A.18}$$

Then define the infeasible estimating term and its average

$$g_t(\lambda) \coloneqq \frac{h_t(\lambda)}{\sigma_t^2(\lambda)} \frac{\partial \vartheta_{t|t-1}(\lambda)}{\partial \lambda}, \qquad G_T(\lambda) \coloneqq \frac{1}{T} \sum_{t=t_0+1}^{T} g_t(\lambda).$$

Let $\hat{\vartheta}_{t|t-1}(\lambda)$ denote the feasible recursion computed from observed data and an arbitrary initialization, and let $\widehat{\vartheta}'_{t|t-1}(\lambda)$ be its derivative recursion. Using Assumption 3, a

feasible $\hat{\sigma}_t^2(\lambda)$ is available. Define

$$\hat{h}_t(\lambda) \coloneqq \psi^{\mathrm{B}}_{\gamma,\xi}\big(Y_t^k - \hat{\vartheta}_{t|t-1}(\lambda)\big), \qquad \hat{g}_t(\lambda) \coloneqq \frac{\hat{h}_t(\lambda)}{\hat{\sigma}_t^2(\lambda)}\,\widehat{\vartheta}'_{t|t-1}(\lambda), \qquad \hat{G}_T(\lambda) \coloneqq \frac{1}{T}\sum_{t=t_0+1}^{T}\hat{g}_t(\lambda),$$

and let $\hat{\lambda}_T$ be any solution to

$$\hat{G}_T(\hat{\lambda}_T) = 0.$$

We show that there exist $K < \infty$ and a nonnegative process U_t such that, for sufficiently large t,

$$\sup_{\lambda\in\Lambda}\|g_t(\lambda) - \hat{g}_t(\lambda)\| \le K\,\rho^t\,U_t \qquad \text{a.s.} \tag{A.19}$$

Use the definition of $g_t(\lambda)$ and add and subtract $\frac{\hat{h}_t(\lambda)}{\hat{\sigma}_t^2(\lambda)}\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}$:

$$\begin{aligned} g_t(\lambda) - \hat{g}_t(\lambda) &= \frac{h_t(\lambda)}{\sigma_t^2(\lambda)}\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda} - \frac{\hat{h}_t(\lambda)}{\hat{\sigma}_t^2(\lambda)}\widehat{\vartheta}'_{t|t-1}(\lambda) \\ &= \underbrace{\left(\frac{h_t(\lambda)}{\sigma_t^2(\lambda)} - \frac{\hat{h}_t(\lambda)}{\hat{\sigma}_t^2(\lambda)}\right)\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}}_{(I)} + \underbrace{\frac{\hat{h}_t(\lambda)}{\hat{\sigma}_t^2(\lambda)}\left(\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda} - \widehat{\vartheta}'_{t|t-1}(\lambda)\right)}_{(II)}. \end{aligned} \tag{A.20}$$

Hence, by the triangle inequality,

$$\|g_t(\lambda) - \hat{g}_t(\lambda)\| \le \|(I)\| + \|(II)\|. \tag{A.21}$$

Start with the first term (I). One can rewrite it to obtain the following,

$$\frac{h_t(\lambda)}{\sigma_t^2(\lambda)} - \frac{\hat{h}_t(\lambda)}{\hat{\sigma}_t^2(\lambda)} = \frac{h_t(\lambda) - \hat{h}_t(\lambda)}{\sigma_t^2(\lambda)} + \hat{h}_t(\lambda)\Big(\frac{1}{\sigma_t^2(\lambda)} - \frac{1}{\hat{\sigma}_t^2(\lambda)}\Big). \tag{A.22}$$

Therefore,

$$\|(I)\| \leq \left[\frac{|h_t(\lambda) - \hat{h}_t(\lambda)|}{\sigma_t^2(\lambda)} + |\hat{h}_t(\lambda)|\Big|\frac{1}{\sigma_t^2(\lambda)} - \frac{1}{\hat{\sigma}_t^2(\lambda)}\Big|\right] \left\|\frac{\partial \vartheta_{t|t-1}(\lambda)}{\partial \lambda}\right\|. \tag{A.23}$$

By the mean value theorem,

$$\begin{aligned}
|h_t(\lambda) - \hat{h}_t(\lambda)| &= \left|\psi^{\mathrm{B}}_{\gamma,\xi}(Y_t^k - \vartheta_{t|t-1}(\lambda)) - \psi^{\mathrm{B}}_{\gamma,\xi}(Y_t^k - \hat{\vartheta}_{t|t-1}(\lambda))\right| \\
&= \left|\frac{\partial \psi^{\mathrm{B}}_{\gamma,\xi}(x_k)}{\partial x_k}\Big|_{x_k=\tilde{x}} \Big(\vartheta_{t|t-1}(\lambda) - \hat{\vartheta}_{t|t-1}(\lambda)\Big)\right| \\
&\leq \sup_{x_k}\left|\frac{\partial \psi^{\mathrm{B}}_{\gamma,\xi}(x_k)}{\partial x_k}\right| \left|\Big(\vartheta_{t|t-1}(\lambda) - \hat{\vartheta}_{t|t-1}(\lambda)\Big)\right| \\
&\leq \frac{1}{\xi^2}\,|\vartheta_{t|t-1}(\lambda) - \hat{\vartheta}_{t|t-1}(\lambda)|.
\end{aligned} \tag{A.24}$$

Where in the last line we use the bound of Barron (2019), $\frac{\partial \psi^{\mathrm{B}}_{\gamma,\xi}(x_k)}{\partial x_k} \leq \frac{1}{\xi^2}$ for $\gamma \leq 2$. First, by (A.18),

$$\frac{1}{\sigma_t^2(\lambda)} \leq \frac{1}{\underline{\sigma}^2}, \qquad \forall \lambda \in \Lambda.$$

Next, Assumption 3(iii) implies that there exists a T_0 such that for all $t \geq T_0$,

$$\sup_{\lambda\in\Lambda} |\hat{\sigma}_t^2(\lambda) - \sigma_t^2(\lambda)| \leq \frac{1}{2}\underline{\sigma}^2.$$

Hence, for $t \geq T_0$ and all λ,

$$\hat{\sigma}_t^2(\lambda) \geq \sigma_t^2(\lambda) - \tfrac{1}{2}\underline{\sigma}^2 \geq \tfrac{1}{2}\underline{\sigma}^2, \qquad \text{so} \qquad \frac{1}{\hat{\sigma}_t^2(\lambda)} \leq \frac{2}{\underline{\sigma}^2}.$$

Therefore, for $t \geq T_0$,

$$\Big|\frac{1}{\sigma_t^2(\lambda)} - \frac{1}{\hat{\sigma}_t^2(\lambda)}\Big| = \frac{|\hat{\sigma}_t^2(\lambda) - \sigma_t^2(\lambda)|}{\sigma_t^2(\lambda)\hat{\sigma}_t^2(\lambda)} \leq \frac{2}{\underline{\sigma}^4}\,|\hat{\sigma}_t^2(\lambda) - \sigma_t^2(\lambda)|. \tag{A.25}$$

Similarly we can bound the second term using the same reasoning to obtain (taking the supremum)

$$\sup_{\lambda\in\Lambda}\|II\| \leq \frac{2}{\underline{\sigma}^2}\Big(\sup_{\lambda\in\Lambda}|\hat{h}_t(\lambda)|\Big)\left(\sup_{\lambda\in\Lambda}\left\|\widehat{\vartheta}'_{t|t-1}(\lambda) - \frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\right\|\right).$$

Combining the two results and taking the supremum, there exists a T_1 such that for all $t \geq T_1$,

$$\begin{aligned}\sup_{\lambda\in\Lambda}\|g_t(\lambda) - \hat{g}_t(\lambda)\| &\leq \sup_{\lambda\in\Lambda}\|(I)\| + \sup_{\lambda\in\Lambda}\|(II)\| \\ &\leq \sup_{\lambda\in\Lambda}\left\|\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\right\|\Big[\frac{1}{\xi^2\underline{\sigma}^2}\sup_{\lambda\in\Lambda}|\hat{\vartheta}_{t|t-1}(\lambda) - \vartheta_{t|t-1}(\lambda)| \\ &\qquad + \frac{2}{\underline{\sigma}^4}\Big(\sup_{\lambda\in\Lambda}|\hat{h}_t(\lambda)|\Big)\Big(\sup_{\lambda\in\Lambda}|\hat{\sigma}_t^2(\lambda) - \sigma_t^2(\lambda)|\Big)\Big] \\ &\quad + \frac{2}{\underline{\sigma}^2}\Big(\sup_{\lambda\in\Lambda}|\hat{h}_t(\lambda)|\Big)\Big(\sup_{\in\Lambda}\left\|\widehat{\vartheta}'_{t|t-1}(\lambda) - \frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\right\|\Big) \quad \text{a.s.}\end{aligned}$$

This expression can be simplified further by noting that

$$|\hat{h}_t(\lambda)| \leq |y_t^k| + |\hat{\vartheta}_{t|t-1}(\lambda)| \leq |y_t^k| + |\vartheta_{t|t-1}(\lambda)| + |\hat{\vartheta}_{t|t-1}(\lambda) - \vartheta_{t|t-1}(\lambda)|,$$

whence

$$\sup_{\lambda\in\Lambda}|\hat{h}_t(\lambda)| \leq \sup_{\lambda\in\Lambda}(|y_t^k| + |\vartheta_{t|t-1}(\lambda)| + |\hat{\vartheta}_{t|t-1}(\lambda) - \vartheta_{t|t-1}(\lambda)|) \leq \sup_{\lambda\in\Lambda}(|y_t^k| + |\vartheta_{t|t-1}(\lambda)| + 1)$$

Furthermore we can use Lemma A.3 and A.4, which hold under Assumption 3, to obtain that there exist finite random constants C_ϑ, $C_{\vartheta'}$, and C_σ such that, for all sufficiently large t,

$$\sup_{\lambda\in\Lambda}|\hat{\vartheta}_{t|t-1}(\lambda) - \vartheta_{t|t-1}(\lambda)| \leq C_\vartheta\rho^t, \qquad \sup_{\lambda\in\Lambda}|\hat{\sigma}_t^2(\lambda) - \sigma_t^2(\lambda)| \leq C_\sigma\rho^t,$$

and

$$\sup_{\lambda\in\Lambda}\|\widehat{\vartheta}'_{t|t-1}(\lambda)-\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\|\leq C_{\vartheta'}\rho^t.$$

Substituting these bounds yields

$$\sup_{\lambda\in\Lambda}\|g_t(\lambda)-\hat{g}_t(\lambda)\|\leq\rho^t\Big[K_1\sup_{\lambda\in\Lambda}\left\|\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\right\|+K_2\sup_{\lambda\in\Lambda}\left\|\frac{\partial\vartheta_{t|t-1}(\lambda)}{\partial\lambda}\right\|\sup_{\lambda\in\Lambda}|\hat{h}_t(\lambda)|+K_3\sup_{\lambda\in\Lambda}|\hat{h}_t(\lambda)|\Big],$$

for some finite constants K_1, K_2, K_3. Using the the bound on $\sup_{\lambda\in\Lambda}|\hat{h}_t(\lambda)|$, this can be rewritten as,

$$\sup_{\lambda\in\Lambda}\|g_t(\lambda)-\hat{g}_t(\lambda)\|\leq K\,\rho^t\,U_t,$$

where

$$U_t\;:=\;\Big(1+\sup_{\lambda\in\Lambda}|\vartheta'_{t|t-1}(\lambda)|\Big)\Big(2+|y_t^k|+\sup_{\lambda\in\Lambda}|\vartheta_{t|t-1}(\lambda)|\Big),\tag{A.26}$$

and $K<\infty$ is a generic constant absorbing K_1, K_2, K_3 and $\underline{\sigma}^2$. This proves (A.19).

By (A.19), there exist $\rho\in(0,1)$ and an a.s. finite random constant $K<\infty$ such that, for all sufficiently large t,

$$\sup_{\lambda\in\Lambda}\|g_t(\lambda)-\hat{g}_t(\lambda)\|\;\leq\;K\,\rho^t\,U_t\qquad\text{a.s.}\tag{A.27}$$

Using Assumption 3(v), $\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\sup_{\lambda\in\Lambda}|\vartheta_{t|t-1}(\lambda)|\right]<\infty$ and $\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\sup_{\lambda\in\Lambda}\left\|\frac{\partial\hat{\vartheta}_{t|t-1}(\lambda)}{\partial\lambda}\right\|\right]<\infty$. From these assumptions and results it follows that, $\mathbb{E}_{\bar{\mathrm{P}}}\log^+U_t<\infty$. Then, by Lemma 2.1 of Straumann and Mikosch (2006),

$$\sum_{t=1}^{\infty}\sup_{\lambda\in\Lambda}\|g_t(\lambda)-\hat{g}_t(\lambda)\|<\infty\quad\text{a.s.}$$

Since $\rho^t\to0$ a.s., K is a finite constant and $\mathbb{R}$ is a Banach space.

Finally, for any T,

$$\sup_{\lambda\in\Lambda}\|G_T(\lambda)-\hat{G}_T(\lambda)\| = \sup_{\lambda\in\Lambda}\Big\|\frac{1}{T}\sum_{t=t_0+1}^{T}\big(g_t(\lambda)-\hat{g}_t(\lambda)\big)\Big\|$$
$$\leq \frac{1}{T}\sum_{t=t_0+1}^{T}\sup_{\lambda\in\Lambda}\|g_t(\lambda)-\hat{g}_t(\lambda)\| \leq \frac{1}{T}\sum_{t=1}^{\infty}\sup_{\lambda\in\Lambda}\|g_t(\lambda)-\hat{g}_t(\lambda)\|,$$

This entails that,

$$\sup_{\lambda\in\Lambda}\|G_T(\lambda)-\hat{G}_T(\lambda)\| = O(T^{-1}) \qquad \text{a.s.} \tag{A.28}$$

Recall,

$$G_T(\lambda) \coloneqq \frac{1}{T}\sum_{t=t_0+1}^{T} g_t(\lambda),$$

and let,

$$G(\lambda) \coloneqq \mathbb{E}_{\bar{\mathrm{P}}}[g_t(\lambda)]\,.$$

Hence, under Assumption 3(v) by the uniform ergodic theorem

$$\sup_{\lambda\in\Lambda}\|G_T(\lambda)-G(\lambda)\| \xrightarrow{a.s.} 0. \tag{A.29}$$

Fix $\lambda\in\Lambda$ with $\lambda\neq\lambda_0$. The condition $\mathbb{E}_{\tilde{\mathrm{P}}_t}h_t(\lambda_0)=0$ (A.17) implies that $\mathbb{E}_{\bar{\mathrm{P}}}[g_t(\lambda_0)] = \mathbb{E}_{\bar{\mathrm{P}}}\left[\mathbb{E}_{\tilde{\mathrm{P}}_t}[g_t(\lambda_0)]\right] = \mathbb{E}_{\bar{\mathrm{P}}}\left[\frac{\partial\vartheta_{t|t-1}(\lambda_0)/\partial\lambda}{\sigma_t^2(\lambda_0)}\mathbb{E}_{\tilde{\mathrm{P}}_t}[h_t(\lambda_0)]\right] = 0$ as $\frac{\partial\vartheta_{t|t-1}(\lambda_0)/\partial\lambda}{\sigma_t^2(\lambda_0)}$ is $\mathcal{F}_{t-1}^{Y,Z}$ measurable. Furthermore by Assumption 3(iv) since $G(\lambda)\neq 0$ and $G(\cdot)$ is continuous, there exists an open neighborhood $V(\lambda)\subset\Lambda$ and a constant $c(\lambda)>0$ such that

$$\inf_{\lambda'\in V(\lambda)}\|G(\lambda')\| \geq 2c(\lambda). \tag{A.30}$$

By the uniform ergodic theorem (A.29), for every $\varepsilon>0$ there exists a (random) integer

$T_\varepsilon < \infty$ such that, for all $T \geq T_\varepsilon$,

$$\sup_{\mu \in \Lambda} \|G_T(\mu) - G(\mu)\| < \varepsilon \qquad \text{a.s.} \tag{A.31}$$

Choosing $\varepsilon = c(\lambda) > 0$, it follows that for all sufficiently large T,

$$\sup_{\mu \in \Lambda} \|G_T(\mu) - G(\mu)\| < c(\lambda) \qquad \text{a.s.}$$

For any $\lambda' \in V(\lambda)$, the reverse triangle inequality yields

$$\|G_T(\lambda')\| \geq \|G(\lambda')\| - \|G_T(\lambda') - G(\lambda')\|.$$

Taking the infimum over $\lambda' \in V(\lambda)$ and using (A.30), we obtain

$$\begin{aligned}
\inf_{\lambda' \in V(\lambda)} \|G_T(\lambda')\| &\geq \inf_{\lambda' \in V(\lambda)} \big[\|G(\lambda')\| - \|G_T(\lambda') - G(\lambda')\|\big] \\
&\geq \inf_{\lambda' \in V(\lambda)} \|G(\lambda')\| - \sup_{\lambda' \in V(\lambda)} \|G_T(\lambda') - G(\lambda')\| \\
&\geq 2c(\lambda) - \sup_{\mu \in \Lambda} \|G_T(\mu) - G(\mu)\|,
\end{aligned}$$

Hence, for all sufficiently large T,

$$\inf_{\lambda' \in V(\lambda)} \|G_T(\lambda')\| \geq c(\lambda) \qquad \text{a.s.}$$

Taking the lower limit concludes that

$$\liminf_{T \to \infty} \inf_{\lambda' \in V(\lambda)} \|G_T(\lambda')\| \geq c(\lambda) > 0 \qquad \text{a.s.} \tag{A.32}$$

Furthermore, continuity of $G(\cdot)$ implies that for every neighborhood $V(\lambda_0)$ of λ_0,

$$\limsup_{T\to\infty} \inf_{\lambda'\in V(\lambda_0)} \|G_T(\lambda')\| = 0 \qquad \text{a.s.} \tag{A.33}$$

We already concluded in (A.29) that

$$\sup_{\lambda\in\Lambda} \|\hat{G}_T(\lambda) - G_T(\lambda)\| \xrightarrow{a.s.} 0.$$

Fix $\lambda \in \Lambda$ with $\lambda \neq \lambda_0$ and let $V(\lambda)$ be the neighborhood from (A.30). By the reverse triangle inequality,

$$\inf_{\lambda'\in V(\lambda)} \|\hat{G}_T(\lambda')\| \geq \inf_{\lambda'\in V(\lambda)} \|G_T(\lambda')\| - \sup_{\mu\in\Lambda} \|\hat{G}_T(\mu) - G_T(\mu)\|. \tag{A.34}$$

Combining this with (A.29) and (A.32) yields

$$\liminf_{T\to\infty} \inf_{\lambda'\in V(\lambda)} \|\hat{G}_T(\lambda')\| \geq c(\lambda) \qquad \text{a.s.} \tag{A.35}$$

Similarly, for any neighborhood $V(\lambda_0)$ of λ_0,

$$\inf_{\lambda'\in V(\lambda_0)} \|\hat{G}_T(\lambda')\| \leq \inf_{\lambda'\in V(\lambda_0)} \|G_T(\lambda')\| + \sup_{\lambda\in\Lambda} \|\hat{G}_T(\lambda) - G_T(\lambda)\|, \tag{A.36}$$

which together with (A.29) and (A.33) imply

$$\limsup_{T\to\infty} \inf_{\lambda'\in V(\lambda_0)} \|\hat{G}_T(\lambda')\| = 0 \qquad \text{a.s.} \tag{A.37}$$

This establishes separation of the minima.

We can now conclude the consistency of $\hat{\lambda}$. Let $\hat{\lambda}_T$ satisfy $\hat{G}_T(\hat{\lambda}_T) = 0$. Suppose, by contradiction, that $\hat{\lambda}_T \not\to \lambda_0$. Since $\hat{\lambda}_T \not\to \lambda_0$, there exist $\varepsilon > 0$ and a subsequence (T') such

that $\|\hat{\lambda}_{T'} - \lambda_0\| \geq \varepsilon$ for all T'. By Assumption 3(ii) Λ is compact, so $(\hat{\lambda}_{T'})$ admits a further convergent subsequence $(\hat{\lambda}_{T_j})_{j\geq 1}$ with limit $\lambda \in \Lambda$ (Bolzano–Weierstrass). By construction $\|\lambda - \lambda_0\| \geq \varepsilon > 0$, so $\lambda \neq \lambda_0$. Let $V(\lambda)$ be any neighborhood of λ. Then $\hat{\lambda}_{T_j} \in V(\lambda) \cap \Lambda$ for all j sufficiently large. Hence,

$$0 = \|\hat{G}_{T_j}(\hat{\lambda}_{T_j})\| \geq \inf_{\mu \in V(\lambda)\cap\Lambda} \|\hat{G}_{T_j}(\mu)\|.$$

But by (A.35), since $V(\lambda) \cap \Lambda$ is a neighborhood of λ in Λ,

$$\liminf_{j\to\infty} \inf_{\mu \in V(\lambda)\cap\Lambda} \|\hat{G}_{T_j}(\mu)\| \geq c(\lambda) > 0 \quad \text{a.s.},$$

a contradiction. Therefore,

$$\hat{\lambda}_T \to \lambda_0 \quad \text{a.s.}$$

□

A.5 Proof of Theorem 4

Proof. The proof follows the proof of Theorem 1 in Blasques et al. (2023) step for step. The only difference is that our estimating function uses the Barron score $h_t = \psi^B_{\gamma,\xi}(Y_t^k - \vartheta_{t|t-1})$ in place of the identity residual $Y_t^k - \vartheta_{t|t-1}$, so that differentiating g_t introduces the factor $\psi^{B\prime}_{\gamma,\xi}(x_t)$ in the term arising from $\partial h_t/\partial\lambda^\top$. Since $|\psi^{B\prime}_{\gamma,\xi}| \leq \xi^{-2}$ for $\gamma \leq 2$ (Barron 2019), this factor is uniformly bounded and enters the domination bounds as a constant, leaving every step unchanged; at $\gamma = 2$ it equals ξ^{-2} and the setting reduces to theirs.

By Theorem 3, $\hat{\lambda}_T \xrightarrow{a.s.} \lambda_0$, with $\lambda_0 \in \mathrm{Int}(\Lambda)$ by Assumption 4(i). Throughout write $\vartheta'_t \coloneqq \partial\vartheta_{t|t-1}(\lambda)/\partial\lambda$, $\vartheta''_t \coloneqq \partial^2\vartheta_{t|t-1}(\lambda)/\partial\lambda\partial\lambda^\top$, $x_t \coloneqq Y_t^k - \vartheta_{t|t-1}(\lambda)$ and $h_t = \psi^B_{\gamma,\xi}(x_t)$.

Differentiating $g_t(\lambda) = \dfrac{h_t(\lambda)}{\sigma_t^2(\lambda)}\,\vartheta_t'(\lambda)$, and using $\partial x_t/\partial\lambda^\top = -\vartheta_t'^\top$,

$$\frac{\partial G_T(\lambda)}{\partial\lambda^\top} = \frac{1}{T}\sum_{t=t_0+1}^{T}\left[\frac{h_t}{\sigma_t^2}\,\vartheta_t'' - \frac{h_t}{\sigma_t^4}\,\vartheta_t'\,\frac{\partial\sigma_t^2}{\partial\lambda^\top} - \frac{\psi_{\gamma,\xi}^{B\prime}(x_t)}{\sigma_t^2}\,\vartheta_t'\,\vartheta_t'^\top\right], \tag{A.38}$$

where $\psi_{\gamma,\xi}^{B\prime}(x_t) \coloneqq \partial\psi_{\gamma,\xi}^{B}(x)/\partial x|_{x=x_t}$ satisfies $|\psi_{\gamma,\xi}^{B\prime}| \le \xi^{-2}$ for $\gamma \le 2$ (Barron 2019), and reduces to the constant ξ^{-2} at $\gamma = 2$. By already given arguments, Lemma A.4 and Assumption 4(iv) show that

$$\sup_{\lambda\in\Lambda}\left\|\frac{\partial G_T(\lambda)}{\partial\lambda^\top} - \frac{\partial\hat{G}_T(\lambda)}{\partial\lambda^\top}\right\| = O(T^{-1}) \quad \text{a.s.} \tag{A.39}$$

Now (A.38), the ergodic theorem and $\mathbb{E}_{t-1}[h_t(\lambda_0)] = 0$ imply

$$\dot{G}_T \coloneqq \frac{\partial G_T(\lambda_0)}{\partial\lambda^\top} \xrightarrow{a.s.} -\mathcal{J}_0, \qquad \mathcal{J}_0 \coloneqq -\,\mathbb{E}_{\bar{\mathrm{P}}}\left[\frac{\partial}{\partial\lambda^\top} g_t(\lambda)\Big|_{\lambda=\lambda_0}\right], \tag{A.40}$$

which by Assumption 4(v) is finite and non–singular.

Doing Taylor expansions of $G_{1T}(\cdot), \ldots, G_{pT}(\cdot)$, where $G_{iT}(\lambda)$ denotes the i-th element of $G_T(\lambda)$, and using (A.28), we have almost surely

$$0_p = \sqrt{T}\,\hat{G}_T(\hat{\lambda}_T) = \sqrt{T}\,G_T(\lambda_0) - \mathcal{J}_T\,\sqrt{T}(\hat{\lambda}_T - \lambda_0) + o(1), \tag{A.41}$$

where the i-th row of $\mathcal{J}_T$ is of the form $-\,\partial G_{iT}(\tilde{\lambda}_{iT})/\partial\lambda^\top$ with $\tilde{\lambda}_{iT}$ a point between $\hat{\lambda}_T$ and λ_0.

The consistency of $\hat{\lambda}_T$ entails that, for T large enough,

$$\left\|\mathcal{J}_T + \dot{G}_T\right\| \le \frac{1}{T}\sum_{t=t_0+1}^{T}\sup_{\lambda\in V(\lambda_0)}\left\|\frac{\partial g_t(\lambda)}{\partial\lambda^\top} - \frac{\partial g_t(\lambda_0)}{\partial\lambda^\top}\right\| \tag{A.42}$$

for any neighbourhood $V(\lambda_0)$ of λ_0. In view of Assumption 4(iii), the floor $\sigma_t^2 \ge \underline{\sigma}^2$

(Assumption 3(iii)), the bound $|\psi^{B\prime}_{\gamma,\xi}| \leq \xi^{-2}$ and the Cauchy–Schwarz inequality, the three terms of (A.38) are dominated:

$$\mathbb{E}_{\bar{\mathrm{P}}} \sup_{\lambda \in V(\lambda_0)} \left\| \frac{h_t}{\sigma_t^2} \vartheta_t'' \right\| \leq \left(\mathbb{E}_{\bar{\mathrm{P}}} \sup_{\lambda \in V(\lambda_0)} \left\| \frac{h_t}{\sigma_t} \right\|^2 \mathbb{E}_{\bar{\mathrm{P}}} \sup_{\lambda \in V(\lambda_0)} \left\| \frac{\vartheta_t''}{\sigma_t} \right\|^2 \right)^{1/2} < \infty,$$

$$\mathbb{E}_{\bar{\mathrm{P}}} \sup_{\lambda \in V(\lambda_0)} \left\| \frac{h_t}{\sigma_t^4} \vartheta_t' \frac{\partial \sigma_t^2}{\partial \lambda^\top} \right\| < \infty, \qquad \mathbb{E}_{\bar{\mathrm{P}}} \sup_{\lambda \in V(\lambda_0)} \left\| \frac{\psi^{B\prime}_{\gamma,\xi}(x_t)}{\sigma_t^2} \vartheta_t' \vartheta_t'^\top \right\| \leq \frac{1}{\xi^2} \mathbb{E}_{\bar{\mathrm{P}}} \sup_{\lambda \in V(\lambda_0)} \left\| \frac{\vartheta_t'}{\sigma_t} \right\|^2 < \infty.$$

In view of (A.38), it follows that, by the ergodic theorem, the right-hand side of (A.42) tends almost surely to

$$\mathbb{E}_{\bar{\mathrm{P}}} \sup_{\lambda \in V(\lambda_0)} \left\| \frac{\partial g_t(\lambda)}{\partial \lambda^\top} - \frac{\partial g_t(\lambda_0)}{\partial \lambda^\top} \right\|.$$

By the dominated convergence theorem this expectation is arbitrarily small when $V(\lambda_0)$ is small. In view of (A.40), it follows that

$$\mathcal{J}_T \xrightarrow{a.s.} \mathcal{J}_0. \tag{A.43}$$

Since (A.39) is uniform in λ, it holds at the feasible intermediate points defining the rows of $\hat{\mathcal{J}}_T$, so $\|\hat{\mathcal{J}}_T - \mathcal{J}_T\| \to 0$ a.s.; with (A.43), the consistency of $\hat{\mathcal{J}}_T$ follows exactly as that of $\mathcal{J}_T$,

$$\hat{\mathcal{J}}_T \xrightarrow{a.s.} \mathcal{J}_0, \tag{A.44}$$

which is non–singular for T large enough, with $\hat{\mathcal{J}}_T^{-1} \xrightarrow{a.s.} \mathcal{J}_0^{-1}$.

Finally, by (A.28), $\sqrt{T}\,(\hat{G}_T(\lambda_0) - G_T(\lambda_0)) \to 0$ a.s.; by (A.17) and the F_{t-1}–measurability of $\vartheta_t'(\lambda_0)$ and $\sigma_t^2(\lambda_0)$, $\{g_t(\lambda_0)\}_{t\in\mathbb{Z}}$ is a strictly stationary and ergodic martingale difference sequence with $\Omega_0 \coloneqq \mathbb{E}_{\bar{\mathrm{P}}}[g_t(\lambda_0) g_t(\lambda_0)^\top]$ finite under Assumption 4(iii). The central limit theorem for stationary ergodic martingale differences (Blasques 2019, p. 24) thus gives

$$\sqrt{T}\, G_T(\lambda_0) = \frac{1}{\sqrt{T}} \sum_{t=t_0+1}^{T} g_t(\lambda_0) \xrightarrow{d} \mathcal{N}(0, \Omega_0). \tag{A.45}$$

Combining (A.41), (A.44) and (A.45) with Slutsky's theorem,

$$\sqrt{T}(\hat{\lambda}_T - \lambda_0) = \mathcal{J}_T^{-1}\sqrt{T}\, G_T(\lambda_0) + o_p(1) \xrightarrow{d} \mathcal{N}\big(0,\, \mathcal{J}_0^{-1}\, \Omega_0\, (\mathcal{J}_0^{-1})^\top\big).$$

□

B Additional proofs and derivations

B.1 Proof of Corollary 1

Proof. The condition on $\mathrm{H}_t(|\check{x}_k| > M)$ together with the positivity assumption on $\mathbb{E}_{\mathrm{H}_t}[g(\check{Y}_t^k, \gamma_1, \gamma_2) \mid |\check{x}_k| > M]$ satisfies the assumptions of Lemma B.1, which yields $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) \geq 0$. The conclusions then follow from Theorem 2. □

Lemma B.1. *Under the setting of Theorem 2, let $\check{x}_k = x_k(\check{Y}_t^k, \vartheta_{t|t-1})$ be the residual map evaluated in $\check{Y}_t^k$ and denote the conditional measure; for any $M > 0$, with $\mathrm{H}_t(|\check{x}_k| > M) > 0$, $\mathrm{H}_t^{>M} := \mathrm{H}_t\,(\cdot \mid |\check{x}_k| > M)$. Suppose there exists $M > 0$ such that $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{>M}}(\phi_1 \| \phi_2) > 0$. Then $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}} \| \phi_2^{\mathrm{B}}) \geq 0$ provided that*

$$\mathrm{H}_t(|\check{x}_k| > M) \;\geq\; \max\left\{0,\; \frac{-\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1 \| \phi_2)}{\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{>M}}(\phi_1 \| \phi_2) - \Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1 \| \phi_2)}\right\},$$

where the right-hand side lies in $[0, 1]$.

B.2 Proof of Lemma B.1

Proof. Define,

$$
\begin{aligned}
g(\check{Y}_t^k, \gamma_1, \gamma_2) \coloneqq \alpha\, \mathbb{E}_{\mathrm{P}_t}\Big[\nabla_\vartheta S(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t^k)\Big]\Big(\psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1}) - \psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})\Big) \\
+ \alpha^2\, \mathbb{E}_{\mathrm{P}_t}\Big[\widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_1))\, \psi_1^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2 \\
- \widetilde{H}_{\mathrm{S}}(Y_t^k, \vartheta_{t|t}(\check{Y}_t^k, \gamma_2))\, \psi_2^{\mathrm{B}}(\check{Y}_t^k, \vartheta_{t|t-1})^2\Big],
\end{aligned}
$$

Now decompose $\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2)\big]$ over the events $\{|\check{x}_k| \le M\}$ and $\{|\check{x}_k| > M\}$

$$
\begin{aligned}
\mathbb{E}_{\mathrm{H}_t}\Big[g(\check{Y}_t^k, \gamma_1, \gamma_2)\Big] = \mathbb{E}_{\mathrm{H}_t}\Big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\Big|\; |\check{x}_k| \le M\Big](1 - \mathrm{H}_t(|\check{x}_k| > M)) \\
+ \mathbb{E}_{\mathrm{H}_t}\Big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\Big|\; |\check{x}_k| > M\Big]\, \mathrm{H}_t(|\check{x}_k| > M).
\end{aligned}
$$

If $\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\big|\; |\check{x}_k| \le M\big] \ge 0$, both conditional expectations are non-negative, so $\mathbb{E}_{\mathrm{H}_t}[g(\check{Y}_t^k, \gamma_1, \gamma_2)] \ge 0$ for all H_t.

When, $\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \mid |\check{x}_k| \le M\big] < 0$ and $\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \mid |\check{x}_k| > M\big] > 0$, the denominator satisfies

$$
\mathbb{E}_{\mathrm{H}_t}\Big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\Big|\; |\check{x}_k| > M\Big] - \mathbb{E}_{\mathrm{H}_t}\Big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\Big|\; |\check{x}_k| \le M\Big] > 0,
$$

and rearranging $(1-p)\,\mathbb{E}_{\mathrm{H}_t}[g(\check{Y}_t^k, \gamma_1, \gamma_2) \mid |\check{x}_k| \le M] + p\,\mathbb{E}_{\mathrm{H}_t}[g(\check{Y}_t^k, \gamma_1, \gamma_2) \mid |\check{x}_k| > M] \ge 0$ for $p \coloneqq \mathrm{H}_t(|\check{x}_k| > M)$ yields the stated bound. The right-hand side equals

$$
\frac{-\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\big|\; |\check{x}_k| \le M\big]}{\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\big|\; |\check{x}_k| > M\big] - \mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\big|\; |\check{x}_k| \le M\big]},
$$

which lies in $(0, 1)$. Recall the conditional measure; for any $M > 0$, with $\mathrm{H}_t(|\check{x}_k| > M) > 0$, $\mathrm{H}_t^{>M} \coloneqq \mathrm{H}_t\,(\cdot \mid |\check{x}_k| > M)$. Gives that $\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\big|\; |\check{x}_k| > M\big] = \Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{>M}}(\phi_1 \| \phi_2)$

and $\mathbb{E}_{\mathrm{H}_t}\big[g(\check{Y}_t^k, \gamma_1, \gamma_2) \;\big|\; |\check{x}_k| \leq M\big] = \Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1 \| \phi_2)$, the right hand side equals

$$\frac{-\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1 \| \phi_2)}{\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{> M}}(\phi_1 \| \phi_2) - \Delta\Delta\mathbb{D}_S^{\mathrm{H}_t^{\leq M}}(\phi_1 \| \phi_2)},$$

□

B.3 Proof of proposition 1

For details on the Von Mises expansion see Appendix C.4

Proof. We treat u_γ as an M–estimator with score function

$$\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k) = \nabla_\vartheta S^{\mathrm{B}}_k(y, \vartheta; \gamma, \xi), \qquad x_k = Y^k - \vartheta,$$

and estimating equation

$$\mathbb{E}_{\mathrm{P}}[\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(Y - u_\gamma(\mathrm{P}))] = 0.$$

By the general Hampel–Huber formula for one–dimensional M–estimators (Hampel et al. 1986, Eq. (2.3.5) in Sec. 2.3a), if $u_{\gamma(F)}$ solves $\int \psi(y, u_\gamma(F))\,\mathrm{dF}(y) = 0$, then

$$\mathrm{IF}(y; u, \mathrm{F}) = \frac{\psi^{\mathrm{B}}_k(y, u_\gamma(\mathrm{F}))}{-\int \psi'^{\mathrm{B}}_{\gamma,\xi}(z, u_\gamma(\mathrm{F}))\,\mathrm{dF}(z)}.$$

we obtain

$$\mathrm{IF}(y; u_\gamma, \mathrm{P}_t) = M_\gamma(\mathrm{P}_t)^{-1}\, \tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}\left(x_k(y, u_\gamma(\mathrm{P}_t))\right),$$

where

$$M_\gamma(\mathrm{P}_t) \coloneqq -\mathbb{E}_{\mathrm{P}_t}\big[\tilde{\psi}'^{\mathrm{B}}_{\gamma,\xi}(x_k(Y, u_\gamma(\mathrm{P}_t))\big].$$

By Hampel's formula, the influence function is

$$\mathrm{IF}(y; u_\gamma, \mathrm{P}_t) = M_\gamma(\mathrm{P}_t)^{-1}\, \tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k(y, u_\gamma(\mathrm{P}_t)),$$

so the influence function is bounded if and only if $\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}$ is bounded, because $M_\gamma(\mathrm{P}_t)$ is a finite nonzero constant under the usual regularity assumptions.

From the explicit derivative bounds of the Barron loss,

$$\left|\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}\left(x_k\right)\right| \;\leq\; \begin{cases} \dfrac{1}{\xi}\left(\dfrac{\gamma-2}{\gamma-1}\right)^{\frac{\gamma-1}{2}} \;\leq\; \dfrac{1}{\xi}, & \gamma \leq 1, \\ \dfrac{|x_k|}{\xi^2}, & \gamma > 1, \end{cases}$$

we see that the score $\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k) = \partial L^{\mathrm{B}}_{\gamma,\xi}/\partial x_k$ is uniformly bounded on $\mathbb{R}$ for $\gamma \leq 1$. For $\gamma > 2$, a lower bound on the score is given by

$$\left|\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}\left(x_k\right)\right| = \frac{|x_k|}{\xi^2}\left(\frac{x_k^2}{\xi^2|\gamma-2|}+1\right)^{\frac{\gamma}{2}-1} > \frac{|x_k|}{\xi^2},$$

which demonstrates that the score is unbounded in that case. For $1 < \gamma < 2$,

$$\left|\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}\left(x_k\right)\right| = \frac{|x_k|}{\xi^2}\left(\frac{x_k^2}{\xi^2|\gamma-2|}+1\right)^{\frac{\gamma}{2}-1} > \frac{|x_k|}{\xi^2}\left(\frac{x_k^2}{\xi^2|\gamma-1|}+x_k^2\right)^{\frac{\gamma}{2}-1} \propto |x_k|^{\frac{\gamma}{2}}, \mathrm{for} |x_k| > 1,$$

so also then the score is unbounded. Finally, for $\gamma = 2$, the Barron loss corresponds to the quadratic loss $\tilde{L}^{\mathrm{B}}_{2,\xi}(x_k) = x_k^2/(2\xi^2)$, with corresponding score satisfying $|\partial \tilde{L}^{\mathrm{B}}_{2,\xi}(x_k)/\partial x_k| = |x_k|/|\xi|$, which is unbounded as well.

Conclude that ψ^{B}_k is bounded if and only if $\gamma \leq 1$, which is equivalent to boundedness of the influence function. Hence u_γ is B–robust at P_t if and only if $\gamma \leq 1$. □

B.4 Consistency verification of the Barron Loss for different values of γ

For this section we focus on the Barron Loss function $L_k^{\mathrm{B}} : \mathcal{Y} \times \Theta \to \mathbb{R}_+ \cup \{\infty\}$, which satisfies $L_k^{\mathrm{B}} = -S_k^{\mathrm{B}}$, to align with the orientation of Gneiting (2011).

Let I be the interval representing the potential range of outcomes of a random variable Y, and let $\mathcal{P}$ be a class of probability distributions in the sample space of I. A loss function $L : I \times I \to [0, \infty)$ is consistent, for a functional $u : \mathcal{P} \to \mathcal{P}_r(I)$ and $\mathrm{F} \to u(\mathrm{F}) \subseteq I$, if

$$\mathbb{E}_{\mathrm{F}}[L(t, Y)] \leq \mathbb{E}_{\mathrm{F}}[L(y, Y)] \tag{B.1}$$

for all probability distributions $\mathrm{F} \in \mathcal{P}$, all $t \in u(\mathrm{F})$, and all $y \in I$.

The loss function L is strictly consistent for the functional u if it is consistent, and equality in the above expression implies that $y \in u(\mathrm{F})$.

The following derivations, show a couple of examples of functionals for which the Barron Loss is strictly consistent. We start by showing consistency of the Barron loss function for the mean if and only if $\gamma = 2$. We specify the error term as $x_k(y_t, \vartheta) = y_t - \vartheta$.

Consider first the case $\gamma = 2$. In this case, the Barron loss reduces to the standard quadratic form

$$\tilde{L}^{\mathrm{B}}_{\gamma,\xi}(x_1(y_t, \vartheta)) = \tfrac{1}{2}\left(\tfrac{y_t - \vartheta}{\xi}\right)^2.$$

Differentiating the expected loss with respect to ϑ yields

$$\frac{\partial}{\partial \vartheta}\mathbb{E}_{\mathrm{P}_t}\left[\tfrac{1}{2}\left(\tfrac{y_t - \vartheta}{\xi}\right)^2\right] = \mathbb{E}_{\mathrm{P}_t}\left[-\tfrac{(y_t - \vartheta)}{\xi^2}\right] = 0,$$

which is solved when $\vartheta = \mathbb{E}_{\mathrm{P}_t}[Y_t]$. Because the loss is convex in ϑ, this solution uniquely minimizes the expected value of the score. Hence, $\vartheta = \mathbb{E}_{\mathrm{P}_t}[Y_t] \in u(\mathrm{P})$ attains the minimal expected score, implying that for $\gamma = 2$ the Barron loss is a strictly $\mathcal{F}$-consistent scoring function for the mean.

Now consider the general case $\gamma \neq \{2, 0, -\infty\}$. In the general case, the Barron loss is defined as

$$\tilde{L}^{\mathrm{B}}_{\gamma,\xi}(x_1(y_t,\vartheta)) = \frac{|\gamma-2|}{\gamma}\left[\left(1+\frac{x_1(y_t,\vartheta)^2}{\xi^2|\gamma-2|}\right)^{\gamma/2}-1\right].$$

Differentiating its expectation with respect to ϑ gives

$$\frac{\partial}{\partial\vartheta}\mathbb{E}_{\mathrm{P}_t}[\tilde{L}^{\mathrm{B}}_{\gamma,\xi}(x_1(y_t,\vartheta))] = \mathbb{E}_{\mathrm{P}_t}\left[-\frac{(y_t-\vartheta)}{\xi^2}\left(1+\frac{(y_t-\vartheta)^2}{\xi^2|\gamma-2|}\right)^{\gamma/2-1}\right] = 0.$$

For these values of γ, the derivative includes a residual-dependent weight term and therefore no longer minimizes at $\vartheta = \mathbb{E}_{\mathrm{P}_t}[Y_t]$ in general.

Next consider the special case $\gamma = 0$. For $\gamma = 0$, the loss simplifies to the logarithmic form

$$\tilde{L}^{\mathrm{B}}_{0,\xi}(x_1(y_t,\vartheta)) = \log\left(\tfrac{1}{2}\left(\tfrac{x_1(y_t,\vartheta)}{\xi}\right)^2+1\right),$$

and differentiation yields

$$\frac{\partial}{\partial\vartheta}\mathbb{E}_{\mathrm{P}_t}[\tilde{L}^{\mathrm{B}}_{0,\xi}(x_1(y_t,\vartheta))] = \mathbb{E}_{\mathrm{P}_t}\left[-\frac{2(y_t-\vartheta)}{(y_t-\vartheta)^2+2\xi^2}\right] = 0.$$

Again, because the weighting term depends on the residual $(y_t - \vartheta)$, the minimizer of the expected loss need not coincide with the mean.

Furthermore consider the limit case where $\gamma = -\infty$. The same arguments hold for the case where $\gamma = -\infty$, in that case, the scoring function is also not minimised at $\vartheta = \mathbb{E}_{\mathrm{P}_t}[Y_t]$. Barron (2019) note that this function instead is minimised at the mode.

We therefore conclude that the Barron loss is strictly consistent for the mean if and only if $\gamma = 2$. For other values of γ, including the limit cases $\gamma = 0$ and $\gamma \to -\infty$, the corresponding score functions target alternative functionals and hence are inconsistent for the mean.

The derivation of strict consistency of the mean, directly translates to strict consistency

of the conditional variance under the assumption that $\mathbb{E}_{\mathrm{P}_t}[Y_t] = 0$. This condition ensures elicitablity of the conditional variance (Fissler and Ziegel 2016). Through reparameterization the problem in terms of $Z_t := Y_t^2$ and considering the induced conditional distribution $Z_t \sim \mathrm{P}_t^{(2)}$. Applying the same arguments as for the mean in this section to Z_t yields strict consistency for the variance functional.

Derivation of consistency of the median is done in a similar manner. Barron (2019) note that for $\gamma = 1$ the loss function is resembles the L^1 (tick) loss that is commonly used for median estimation. The correspondence is approximate for finite positive ξ because the loss function does not exactly reduce to L^1 loss for $\gamma = 1$, but rather a smoothed version of the L^1 loss (Charbonnier loss). The loss function reduces to

$$L^{\mathrm{B}}_{1,\xi}(y_t, \vartheta) = \sqrt{\left(\frac{y_t - \vartheta}{\xi}\right)^2 + 1} - 1.$$

Barron (2019) notes that the loss function is approximately an L^1 loss function for ξ small. Indeed, one finds

$$\tilde{L}^{\mathrm{B}}_{1,\xi}(y_t - \vartheta) \approx \frac{|y_t - \vartheta|}{\xi} - 1$$

for values of ξ that are much smaller than $y_t - \vartheta$. To obtain exact correspondence, it is useful to consider a linearly rescaled version of the Barron loss (this rescaling preserving the location of the minimum of the loss function) for $\gamma = 1$ in the limit as $\xi \downarrow 0$:

$$|\xi| \left(1 + \tilde{L}^{\mathrm{B}}_{1,\xi}\left(y_t - \vartheta\right)\right) = \sqrt{|y_t - \vartheta|^2 + \xi^2} \underset{\xi \downarrow 0}{\longrightarrow} |y_t - \vartheta|.$$

We next show that this function in expectation minimises at $\vartheta = \mathrm{median}(y_t)$ if the median is unique and y_t follows a distribution with pdf $p_t(\cdot)$. Taking expectations one finds

$$\mathbb{E}_{\mathrm{F}}\left[|y_t - \vartheta|\right] = \left(\int_{-\infty}^{\vartheta} (y_t - \vartheta) p(y_t)\, \mathrm{d}y_t - \int_{\vartheta}^{\infty} (y_t - \vartheta) p(y_t)\, \mathrm{d}y_t\right).$$

Now, taking the derivative results in

$$
\begin{aligned}
&\frac{\partial}{\partial\vartheta}\left(\int_{-\infty}^{\vartheta}(y_t-\vartheta)p(y_t)\,\mathrm{d}y_t-\int_{\vartheta}^{\infty}(y_t-\vartheta)p(y_t)\,\mathrm{d}y_t\right)\\
=&\left(\frac{\partial}{\partial\vartheta}\int_{-\infty}^{\vartheta}(y_t-\vartheta)p(y_t)\,\mathrm{d}y_t-\frac{\partial}{\partial\vartheta}\int_{\vartheta}^{\infty}(y_t-\vartheta)p(y_t)\,\mathrm{d}y_t\right)\\
=&\left(\left[\int_{-\infty}^{\vartheta}\frac{\partial}{\partial\vartheta}(y_t-\vartheta)p(y_t)\,\mathrm{d}y_t+(\vartheta-\vartheta)p(\vartheta)\right]\right.\\
&\qquad\left.-\left[-(\vartheta-\vartheta)p(\vartheta)+\int_{\vartheta}^{\infty}\frac{\partial}{\partial\vartheta}(y_t-\vartheta)p(y_t)\,\mathrm{d}y_t\right]\right)\\
=&\left(\int_{-\infty}^{\vartheta}-p(y_t)\,\mathrm{d}y_t+\int_{\vartheta}^{\infty}p(y_t)\,\mathrm{d}y_t\right)\\
=&(1-2P(\vartheta)).
\end{aligned}
$$

In the third line Leibniz's rule has been applied because the integration bound depends on the variable being differentiated over. Setting the last line to zero to find the minimum reveals that $P(\vartheta)=0.5$, implying that ϑ is the median of Y_t. Therefore the function is consistent for the median for $\gamma=1$ in the limit as $\xi\downarrow 0$.

For the limiting case that $\gamma=-\infty$, the Barron loss is shown to be consistent for the (local) mode. For this purpose, let $x_1(y,\vartheta)=y-\vartheta$ for the location model. For $\gamma=-\infty$ the Barron loss used in the QSD update is

$$
\tilde{L}^{\mathrm{B}}_{-\infty,\xi}(x_1)=1-\exp\Big(-\tfrac{1}{2}(x_k/\xi)^2\Big),
$$

and its derivative w.r.t. the time-varying parameter ϑ is

$$
\psi^{\mathrm{B}}_{-\infty,\xi}(y,\vartheta)=\frac{\partial L^{\mathrm{B}}_{-\infty,\xi}}{\partial\vartheta}=\frac{\partial\tilde{L}^{\mathrm{B}}_{-\infty,\xi}}{\partial x_k}\cdot\frac{\partial x_k}{\partial\vartheta}=-\frac{(y-\vartheta)}{\xi^2}\exp\Big(-\tfrac{1}{2}\big(\tfrac{y-\vartheta}{\xi}\big)^2\Big).
$$

These formulas match the definitions in Table 1.

We aim to minimize the expected loss

$$\min_{\vartheta} \int_{\mathcal{Y}} p(y) L^{\mathrm{B}}_{\gamma,\xi}(y-\vartheta,\xi)\mathrm{d}y.$$

This has first order condition (assuming the derivative and integral can be interchanged)

$$0 = \int_{\mathcal{Y}} p(y) \frac{(y-\vartheta)}{\xi^2} \exp\left(-\frac{1}{2}\frac{(y-\vartheta)^2}{\xi^2}\right) \mathrm{d}y.$$

Rearranging terms gives

$$\int_{\mathcal{Y}} p(y) \frac{y}{\xi^2} \exp\left(-\frac{1}{2}\frac{(y-\vartheta)^2}{\xi^2}\right) \mathrm{d}y = \vartheta \int_{\mathcal{Y}} \frac{p(y)}{\xi^2} \exp\left(-\frac{1}{2}\frac{(y-\vartheta)^2}{\xi^2}\right) \mathrm{d}y,$$

which results in

$$\vartheta = \frac{\int_{\mathcal{Y}} p(y) \frac{y}{\xi^2} \exp\left(-\frac{1}{2}\frac{(y-\vartheta)^2}{\xi^2}\right) \mathrm{d}y}{\int_{\mathcal{Y}} \frac{p(y)}{\xi^2} \exp\left(-\frac{1}{2}\frac{(y-\vartheta)^2}{\xi^2}\right) \mathrm{d}y} =: u(\vartheta).$$

This is a fixed-point equation and leads to the iterative scheme

$$\vartheta^{(k+1)} = T(\vartheta^{(k)}) = \frac{\int_{\mathcal{Y}} p(y)\, y\, \exp\big(-\frac{1}{2}((y-\vartheta^{(k)})/\xi)^2\big)\, \mathrm{d}y}{\int_{\mathcal{Y}} p(y)\, \exp\big(-\frac{1}{2}((y-\vartheta^{(k)})/\xi)^2\big)\, \mathrm{d}y}. \tag{B.2}$$

Equation (B.2) defines the mean-shift iteration, a fixed-point mapping $\vartheta^{(k+1)} = T(\vartheta^{(k)})$ that moves ϑ in the direction of the weighted average of the data under Gaussian weights centered at the current estimate (Comaniciu and Meer 2002). The update step

$$\vartheta^{(k+1)} - \vartheta^{(k)} = \frac{\int p(y)\,(y-\vartheta^{(k)})\, \exp\big(-\frac{1}{2}((y-\vartheta^{(k)})/\xi)^2\big)\mathrm{d}y}{\int p(y)\, \exp\big(-\frac{1}{2}((y-\vartheta^{(k)})/\xi)^2\big)\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}\vartheta} p_\xi(\vartheta)$$

is proportional to the gradient of the population level version of a Gaussian kernel estimator.

Here the Gaussian kernel estimator is given by

$$p_\xi(\vartheta) = \int_{\mathcal{Y}} p(y)\, \frac{1}{\sqrt{2\pi}\xi} \exp\Big(-\tfrac{1}{2}\big(\tfrac{y-\vartheta}{\xi}\big)^2 \Big)\, \mathrm{d}y.$$

with gradient

$$\frac{\mathrm{d}}{\mathrm{d}\vartheta} p_\xi(\vartheta) = \int_{\mathcal{Y}} p(y)\, \frac{(y-\vartheta)}{\xi^2} \exp\Big(-\tfrac{1}{2}\big(\tfrac{y-\vartheta}{\xi}\big)^2 \Big) \mathrm{d}y,$$

so the first-order condition $\mathbb{E}_{\mathrm{P}_t}[\psi^{\mathrm{B}}_{-\infty,\xi}(Y,\vartheta)] = 0$ is equivalent to $\nabla_\vartheta p_\xi(\vartheta) = 0$. Consequently, every fixed point of (B.2) is a stationary point of the smoothed density $p_\xi(\vartheta)$. When p_ξ has isolated maxima, each such point corresponds to a (local) mode.

We hence conclude that for $\gamma = -\infty$, the Barron loss yields an expected score whose zeros coincide with the stationary points of the Gaussian–smoothed density $p_\xi(\vartheta)$. Minimizing the expected loss therefore selects the local mode functional

$$u(\mathrm{P}) = \arg\max_\vartheta p_\xi(\vartheta),$$

establishing (local) mode consistency of the Barron loss in this limit. The associated fixed-point iteration (B.2) recovers the classical mean-shift procedure, and the QSD update thus uses the local-mode seeking mechanism of Van den Boomgaard and Van de Weijer (2002).

C Further details

Properties of the Barron function

To support the use of the Barron (2019) loss, we briefly review its key properties. Reparameterization gives $\psi^{\mathrm{B}}_k(y,\vartheta;\gamma,\xi) \equiv \tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))$, where

$$\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k) \coloneqq \frac{x_k}{\xi^2}\left(1 + \frac{x_k^2}{|\gamma-2|\,\xi^2}\right)^{\gamma/2-1}.$$

Figure 1 shows the Barron loss function and its first order derivative for different values of γ. Specific values of γ reproduce several existing loss functions: L^2 loss ($\gamma = 2$), Charbonnier loss ($\gamma = 1$), Cauchy loss ($\gamma = 0$), Geman-McClure loss ($\gamma = -2$) and Welsch loss ($\gamma = -\infty$). From Figure 1, it can be seen that lower values of γ decrease the effect of large values of x_k on the loss. The same holds for the effect of large values of x_k on the derivative of the loss function, which is relevant because this is used in the update equation for the time-varying parameters.

Barron (2019) mentions some additional properties of the loss function. For instance, the loss function is smooth (C^∞) with respect to x_k, γ and ξ. Also, the loss function has bounded gradient,

$$\left|\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))\right| \leq \begin{cases} \frac{1}{\xi}\left(\frac{\gamma-2}{\gamma-1}\right)^{\left(\frac{\gamma-1}{2}\right)} \leq \frac{1}{\xi} & \text{if } \gamma \leq 1, \\ \\ \frac{|x_k|}{\xi^2} & \text{if } \gamma \leq 2, \end{cases} \tag{C.1}$$

and

$$\frac{\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))}{\partial x_k} \leq \frac{1}{\xi^2}, \quad \text{if } \gamma \leq 2. \tag{C.2}$$

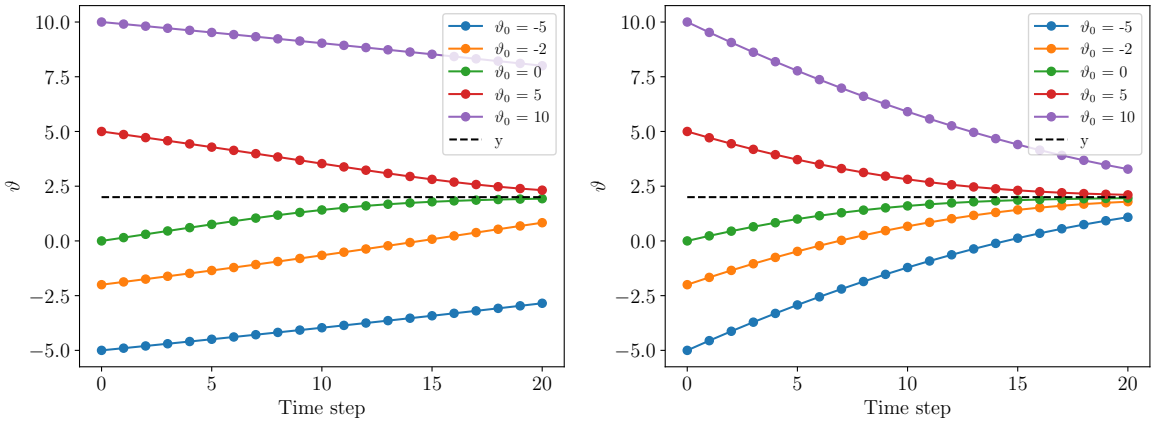
These properties imply that for values of γ less than 1, the derivatives are bounded, preventing gradient explosion. For $\gamma \leq 2$ the gradient is bounded by a function that is linearly dependent on x_k. The loss function also has the shape parameter $\xi > 0$, which determines the shape of the quadratic bowl near $x_k = 0$.

Barron (2019) explains that the first order derivative gives some intuition about how γ affects the behavior of the loss function when it is being minimized by gradient descent. For a given value of γ, the derivative is approximately linear when $|x_k| < \xi$. This means that the effect of the residual is linearly proportional to its magnitude. If $\gamma = 2$, the derivative stays linearly proportional to its magnitude. If γ is decreased to 1, the derivative tends to $1/\xi$ as $|x_k|$ increases. For $\gamma < 1$, the derivative is redescending when $|x_k| > \xi$. The

parameters γ and ξ offer great flexibility to tune the trade-off between responsiveness and robustness when used as the loss function in the QSD framework.

C.1 Illustration of the Convergence Dynamics

Figure C.1: Convergence Dynamics under Different Values of γ



(a) $\gamma = 0.5$ (b) $\gamma = 1.5$

NOTE: Convergence of the estimate $\vartheta_{t|t-1}$ under the recursion (4) when observing $y_t = 2$ for 20 time steps. The different curves correspond to different starting values of ϑ_0. In this illustration, ξ is set to 1. Figure (a), $\gamma = 0.5$ and figure (b), $\gamma = 1.5$ demonstrate that larger γ, yields larger update steps and faster convergence, while smaller γ produces more conservative adjustments.

Figure C.1 illustrates the trajectories of the estimate $\vartheta_{t|t-1}$ under the recursion (4) in a situation where observations with the same value $y_t = 2$ become available for 20 consecutive time steps. Assuming a location model, we see that $\vartheta_{t|t-1}$ converges to the true value $y = 2$, as expected. The effect of γ is immediately visible, a larger value of γ (C.1b) produces larger update steps, whereas a smaller γ yields more conservative adjustments. This highlights the role of γ in controlling the sensitivity of the loss to large residuals, thereby providing robustness against outliers.

C.2 Example illustrating Theorem 1

Example C.1. *Consider the following setting, let Y_t be conditionally Student's-t distributed with 4 degrees of freedom, i.e., $Y_t \mid \mathcal{F}_{t-1} \sim t_4(0, 1)$, take $k = 1$ and fix $\vartheta_{t|t-1} = 1$, fur-*

thermore take $\xi = 1$, $\gamma_1 = 1$, $\gamma_2 = 2$. In this setting the non-robust update ($\gamma_2 = 2$), coincides with an ARMA model. Evaluate the updates using the quadratic scoring rule $S(F_\vartheta, y) := -(y - \vartheta)^2$. Since $\frac{\partial^2 S}{\partial \vartheta^2} = -2$ is a constant, the Taylor expansion in Theorem 1 is exact with $c = 2$.

By symmetry of the Student's-t distribution around zero, $\mathbb{E}_{\mathrm{P}_t}[\nabla_\vartheta S(F_1, Y_t)] = -2$ and $\mathbb{E}_{\mathrm{P}_t}[\psi_i^{\mathrm{B}}(\check{Y}_t, 1)] < 0$ for both $i \in \{1, 2\}$, so Assumption 1(v) holds. For $\gamma_2 = 2$, $\psi_2^{\mathrm{B}}(\check{Y}_t, 1) = \check{Y}_t - 1$ gives $\mathbb{E}_{\mathrm{P}_t}\big[\psi_2^{\mathrm{B}}\big] = -1$ and $\mathbb{E}_{\mathrm{P}_t}\big[(\psi_2^{\mathrm{B}})^2\big] = \frac{4}{2} + 1 = 3$ exactly. For $\gamma_1 = 1$, numerical integration under t_4 gives $\mathbb{E}_{\mathrm{P}_t}\big[\psi_1^{\mathrm{B}}\big] \approx -0.476$ and $\mathbb{E}_{\mathrm{P}_t}\big[(\psi_1^{\mathrm{B}})^2\big] \approx 0.510$. Using the results of Theorem 1, the upper and lower bound of the interval are given by

$$\bar{\alpha}_t(\gamma_i) = \frac{2}{c} \cdot \frac{\mathbb{E}_{\mathrm{P}_t}\big[\nabla_\vartheta S(F_1, Y_t)\big]\, \mathbb{E}_{\mathrm{P}_t}\big[\psi_i^{\mathrm{B}}\big]}{\mathbb{E}_{\mathrm{P}_t}\big[(\psi_i^{\mathrm{B}})^2\big]}, \quad \underline{\alpha}_t = \frac{\mathbb{E}_{\mathrm{P}_t}\big[\nabla_\vartheta S(F_1, Y_t)\big](\mathbb{E}_{\mathrm{P}_t}\big[\psi_2^{\mathrm{B}}\big] - \mathbb{E}_{\mathrm{P}_t}\big[\psi_1^{\mathrm{B}}\big])}{\mathbb{E}_{\mathrm{P}_t}\big[(\psi_2^{\mathrm{B}})^2\big] - \mathbb{E}_{\mathrm{P}_t}\big[(\psi_1^{\mathrm{B}})^2\big]}.$$

Direct computation results in, $\bar{\alpha}_t(\gamma_2) = \frac{2}{3}$, $\bar{\alpha}_t(\gamma_1) \approx 1.866$ and $\underline{\alpha}_t \approx 0.421$.

Hence the interval $[\underline{\alpha}_t, \bar{\alpha}_t) = [0.421, \frac{2}{3})$ is non-empty. For any α in this interval, both updates are individually divergence-reducing and ϕ_1^{B} achieves a weakly larger expected divergence reduction than ϕ_2^{B}.

Two remarks are in place: first note that this example illustrates only the update step $\vartheta_{t|t}$ and excludes the prediction step $\vartheta_{t+1|t}$; the interval $[\underline{\alpha}_t, \bar{\alpha}_t) = [0.421, \frac{2}{3})$ should therefore not be compared directly to learning rates calibrated for a full recursive filter, where α compounds over time through repeated interaction with the persistence parameter β. Second, the interval $[\underline{\alpha}_t, \bar{\alpha}_t)$ over which Theorem 1 establishes formal P_t-contractive dominance arises as the intersection of two separate conditions: the divergence-reduction comparison $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^B \,\|\, \phi_2^B) \geq 0$ and the joint PRADA admissibility region. In the present example the comparison itself holds for every $\alpha \geq \underline{\alpha}_t = 0.421$, but the non-robust update ceases to be PRADA at $\bar{\alpha}_t(\gamma_2) = \frac{2}{3}$, while the robust update remains PRADA up to $\bar{\alpha}_t(\gamma_1) \approx 1.866$. On the resulting gap $\bar{\alpha}_t(\gamma_2) \leq \alpha \leq \bar{\alpha}_t(\gamma_1)$, the two thresholds separate and ϕ_1^B cannot be

said to P_t*-contractively dominate* ϕ_2^B *in the sense of Definition 4: with* $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_2^B) \leq 0$*, the non-robust update no longer contracts toward* P_t*, so the comparison is between a divergence-reducing and a divergence-increasing update rather than between two contractions of differing magnitudes. Nevertheless,* $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^B \,\|\, \phi_2^B) > 0$ *throughout this interval and* ϕ_1^B *remains PRADA, so the robust update preserves its divergence-reducing behaviour over a strictly larger range of* α *than the non-robust one. This asymmetry motivates retaining the joint PRADA cap* $\bar{\alpha}_t = \min_{i\in\{1,2\}} \bar{\alpha}_t(\gamma_i)$ *in Theorem 1: outside this cap a positive sign of* $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}$ *reflects the failure of* ϕ_2^B *as a contraction rather than a genuine relative improvement of* ϕ_1^B *over a competing contraction.*

C.3 Example illustrating Theorem 2

Example C.2. *Consider the contaminated setting of Theorem 2, let* Y_t *be conditionally standard normal under the clean measure,* $\mathrm{P}_t = \mathcal{N}(0,1)$*, and let the contaminating distribution be a scale-3 Student's-*t_4*,* $\mathrm{H}_t = 3 \cdot t_4$*, where* $\mathrm{P}_t^\varepsilon = (1-\varepsilon)\mathrm{P}_t + \varepsilon\mathrm{H}_t$ (8)*. Consider the situation where* $k = 1$*, fix* $\vartheta_{t|t-1} = 1$*,* $\xi = 1$*,* $\gamma_1 = 1$*,* $\gamma_2 = 2$*, and* $S(\mathrm{F}_\vartheta, y) = -(y-\vartheta)^2$*, so that* $c = 2$ *and* $\mathbb{E}_{\mathrm{P}_t}[\nabla_\vartheta S(\mathrm{F}_1, Y_t^k)] = -2$*. In this situation the non-robust update (*$\gamma_2 = 2$*), coincides with ARMA (*$k = 1$*) model.*

For $\gamma_2 = 2$*,* $\psi_2^{\mathrm{B}}(\check{Y}_t, 1) = \check{Y}_t - 1$ *gives* $\mathbb{E}_{\mathrm{P}_t}\big[\psi_2^{\mathrm{B}}\big] = \mathbb{E}_{\mathrm{H}_t}\big[\psi_2^{\mathrm{B}}\big] = -1$*,* $\mathbb{E}_{\mathrm{P}_t}\big[(\psi_2^{\mathrm{B}})^2\big] = 2$*, and* $\mathbb{E}_{\mathrm{H}_t}\big[(\psi_2^{\mathrm{B}})^2\big] = 19$ *exactly. For* $\gamma_1 = 1$*, numerical integration gives* $\mathbb{E}_{\mathrm{P}_t}\big[\psi_1^{\mathrm{B}}\big] \approx -0.514$*,* $\mathbb{E}_{\mathrm{P}_t}\big[(\psi_1^{\mathrm{B}})^2\big] \approx 0.479$*,* $\mathbb{E}_{\mathrm{H}_t}\big[\psi_1^{\mathrm{B}}\big] \approx -0.224$*,* $\mathbb{E}_{\mathrm{H}_t}\big[(\psi_1^{\mathrm{B}})^2\big] \approx 0.710$*.*

By the mixture decomposition,

$$\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon}(\phi_i^{\mathrm{B}}) = (1-\varepsilon)\,\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_i^{\mathrm{B}}) + \varepsilon\,\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_i^{\mathrm{B}}), \qquad \Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^\varepsilon} = (1-\varepsilon)\,\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t} + \varepsilon\,\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t}.$$

Fix $\alpha = 0.5$*, direct computation yields* $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1^{\mathrm{B}}) \approx 0.394$*,* $\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_2^{\mathrm{B}}) = 0.500$*,* $\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_1^{\mathrm{B}}) \approx$

0.046, $\Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_2^{\mathrm{B}}) = -3.750$, $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t} \approx -0.106$, *and* $\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t} \approx 3.796$. *Two thresholds arise:*

$$\underline{\varepsilon} = \frac{-\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}}{\Delta\Delta\mathbb{D}_S^{\mathrm{H}_t} - \Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}} \approx 0.027,$$
$$\bar{\varepsilon} = \frac{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_2^{\mathrm{B}})}{\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_2^{\mathrm{B}}) - \Delta\mathbb{D}_S^{\mathrm{H}_t}(\phi_2^{\mathrm{B}})} \approx 0.118.$$

Hence the admissible range is $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}] \approx [0.027, 0.118]$. *Below* $\underline{\varepsilon}$*, the variance penalty under contamination is too small to flip dominance and the non-robust update remains preferable. Above* $\bar{\varepsilon}$*, the contamination is severe enough that* ϕ_2^{B} *overshoots and ceases to be PRADA on the mixture.*

C.4 Von Mises expansion

As noted by Hampel et al. (1986), the functional $u : \mathcal{P} \to \Theta$ admits a first–order von Mises expansion around P_t: if a distribution G is "close" to P_t, then

$$u(\mathrm{G}) = u(\mathrm{P}_t) + \int \mathrm{IF}(y; u, \mathrm{P}_t)\,\big(\mathrm{G} - \mathrm{P}_t\big)(\mathrm{d}y) + o\big(\|\mathrm{G} - \mathrm{P}_t\|_{\mathrm{TV}}\big). \tag{C.3}$$

With $\|G - P_0\|_{\mathrm{TV}} \coloneqq \sup_{|f|\le 1} \big|\int f\, d(G - P_0)\big|$.

Specializing (C.3) to the contamination model $\mathrm{G} = \mathrm{P}_t^{\varepsilon} = (1-\varepsilon)\mathrm{P}_t + \varepsilon\mathrm{H}_t$ yields

$$\mathrm{d}(\mathrm{G} - \mathrm{P}_t) = \mathrm{d}\mathrm{P}_t^{\varepsilon} - \mathrm{d}\mathrm{P}_t = \varepsilon\,\mathrm{d}(\mathrm{H}_t - \mathrm{P}_t),$$

and hence, to first order in ε,

$$u(\mathrm{P}_t^{\varepsilon}) = u(\mathrm{P}_t) + \varepsilon \int \mathrm{IF}(y; u, \mathrm{P}_t)\,\big(\mathrm{H}_t - \mathrm{P}_t\big)(\mathrm{d}y) + o(\varepsilon). \tag{C.4}$$

In the particularly important case $\mathrm{H}_t = \delta_y$, we obtain

$$u\big((1-\varepsilon)\mathrm{P}_t + \varepsilon\delta_y\big) = u(\mathrm{P}_t) + \varepsilon\,\mathrm{IF}(y; u, \mathrm{P}_t) + o(\varepsilon),$$

so that the influence function $\mathrm{IF}(y; u, \mathrm{P}_t)$ is the coefficient of the linear term in the contamination level ε. This representation directly motivates the notion of B-robustness: boundedness of the influence function ensures that infinitesimal contamination cannot induce arbitrarily large changes in the functional u.

C.5 Influence function of the Barron functional

We work in a static setting with $Y \sim \mathrm{P}$ and residual $x_k = y^k - \vartheta$. For a given shape parameter γ and scale parameter $\xi > 0$, the Barron loss given in (1) with associated score for $\gamma \notin \{2, 0, -\infty\}$ is given by

$$\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k) = \frac{x_k}{\xi^2}\left(1 + \frac{x_k^2}{|\gamma-2|\,\xi^2}\right)^{\gamma/2-1}.$$

The corresponding functional is the population minimizer

$$u_\gamma(\mathrm{P}) \coloneqq \arg\max_{\vartheta\in\Theta} \mathbb{E}_{\mathrm{P}}\left[S^{\mathrm{B}}_k(Y, \vartheta; \gamma, \xi)\right],$$

which satisfies the estimating equation

$$\mathbb{E}_{\mathrm{P}}\left[\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k(Y, u_\gamma(\mathrm{P}))\right] = 0.$$

By M–estimator theory (see, e.g., Hampel et al. 1986, Eq. (2.3.5)), the influence function of u_γ at P is

$$\mathrm{IF}(x_k; u_\gamma, \mathrm{P}) = M_\gamma(\mathrm{P})^{-1}\,\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}\big(x_k(Y, u_\gamma(\mathrm{P})\big), \tag{C.5}$$

where the sensitivity term is the scalar

$$M_\gamma(\mathrm{P}) = \mathbb{E}_\mathrm{P}\left[\tilde{\psi}'^{\mathrm{B}}_{\gamma,\xi}(Y - u_\gamma(\mathrm{P}))\right]. \tag{C.6}$$

Hence the robustness of u_γ is entirely determined by the growth of the score ψ^{B}_k. Substituting $x_k = y - u_\gamma(\mathrm{P})$ gives

$$\tilde{\psi}^{\mathrm{B}}_k(x_k) = \frac{x_k}{\xi^2}\left(1 + \frac{x_k^2}{|\gamma - 2|\,\xi^2}\right)^{\gamma/2-1}.$$

For $\gamma = 2$, the score simplifies to $\psi_2(x_k) = x_k/\xi^2$, and therefore grows linearly in x_k, resulting in an unbounded influence function.

For $\gamma < 2$, the exponent $\gamma/2 - 1$ is negative, so the factor $\left(1 + x_k^2/(|\gamma - 2|\xi^2)\right)^{\gamma/2-1}$ downweights large residuals. Using the explicit bound

$$\left|\tilde{\psi}^{\mathrm{B}}_{\gamma,\xi}(x_k)\right| \;\le\; \frac{1}{\xi} \qquad \text{whenever } \gamma \le 1,$$

we see that the score is uniformly bounded on $\mathbb{R}$ for $\gamma \le 1$, and increases strictly slower than linearly for $1 < \gamma < 2$.

Combining this bound with (10) shows that the Barron functional has a bounded influence function if and only if $\gamma \le 1$. Thus, in the sense of Hampel et al. (1986), u_γ is *B–robust* for $\gamma \le 1$, while the quadratic case $\gamma = 2$ is not robust. The tuning parameter γ therefore acts as a continuous robustness index, with smaller values providing stronger protection against the effect of outlying observations.

Example C.3. *Influence of contamination under a Gaussian DGP*

To illustrate the expansion in (C.3)*, consider a Gaussian location model* $Y = \mu_0 + \sigma Z$ *with* $Z \sim \mathcal{N}(0,1)$*, and set* $\xi = 1$ *for simplicity. The Barron location functional* $u_\gamma(\mathrm{P})$ *is*

defined by the estimating equation

$$\mathbb{E}_{\mathrm{P}}\left[\psi_k^{\mathrm{B}}(Y,\vartheta;\gamma,\xi)\right]=0, \qquad \psi_k^{\mathrm{B}}(y,u_\gamma;\gamma,\xi)=(y-\vartheta)\left(1+\frac{(y-\vartheta)^2}{|\gamma-2|}\right)^{\gamma/2-1}.$$

Since ψ_k^{B} is odd and $Y-\mu_0$ is symmetrically distributed, Fisher consistency holds and $u_\gamma(\mathrm{P}_t)=\mu_0$ for all $\gamma<2$.

Let $\mathrm{P}_t^\varepsilon=(1-\varepsilon)\mathrm{P}_t+\varepsilon\mathrm{H}_t$ denote a contaminated distribution. Applying the von Mises expansion gives

$$u_\gamma(\mathrm{P}_t^\varepsilon)=\mu_0+\varepsilon\,\frac{\int\psi_k^{\mathrm{B}}(y,\mu_0;\gamma,\xi)\,\mathrm{d}(\mathrm{H}_t-\mathrm{P}_t)(y)}{\mathbb{E}_{\mathrm{P}_t}[\psi_k'^{\mathrm{B}}(y,\mu_0;\gamma,\xi)]}+o(\varepsilon).$$

Using the estimating equation $\mathbb{E}_{\mathrm{P}_t}[\psi_k^{\mathrm{B}}(Y_t,\mu_0;\gamma,\xi)]=0$, the numerator can be seen to simplify to $\mathbb{E}_{\mathrm{H}_t}[\psi_k^{\mathrm{B}}(Y_t,\mu_0;\gamma,\xi)]$, yielding

$$u_\gamma(\mathrm{P}_t^\varepsilon)=\mu_0+\varepsilon\,\frac{\mathbb{E}_{\mathrm{H}_t}[\psi_k^{\mathrm{B}}(Y_t,\mu_0;\gamma,\xi)]}{\mathbb{E}_{\mathrm{P}_t}[\psi_k'^{\mathrm{B}}(Y_t,\mu_0;\gamma,\xi)]}+o(\varepsilon).$$

For $\gamma=2$, the Barron loss reduces to the quadratic loss and the score simplifies to $\tilde{\psi}_{\gamma,\xi}^{\mathrm{B}}(x_k)=x_k$. Consequently, the influence function is linear in x_k and therefore unbounded, coinciding with the classical influence function of the sample mean.

For $1<\gamma<2$, the score $\tilde{\psi}_{\gamma,\xi}^{\mathrm{B}}(x_k(y,u_\gamma))$ grows slower than linearly as $|x_k(y,u_\gamma)|\to\infty$, which attenuates but does not eliminate the impact of extreme observations. In this case, the influence function remains unbounded, though its growth rate is reduced relative to the quadratic loss.

For $\gamma\leq 1$, the score is uniformly bounded (and redescending for $\gamma<1$), implying that the influence function

$$\mathrm{IF}(x_k;u_\gamma,\mathrm{P}_t)=M_\gamma(\mathrm{P}_t)^{-1}\psi_k^{\mathrm{B}}(y,u_\gamma;\gamma,\xi)$$

is bounded on $\mathbb{R}$. Hence full B–robustness is attained exactly when $\gamma \leq 1$.

C.6 Asymptotic linearity and the Cramér–Rao lower bound

Let $X_1, \ldots, X_n$ be a strictly stationary and ergodic sequence with marginal distribution P_t, and denote by

$$\mathrm{P}_n(x_k) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x_k\}$$

the empirical cumulative distribution function. By the Glivenko–Cantelli theorem for strictly stationary ergodic sequences (Tucker 1959), $\|\mathrm{P}_n - \mathrm{P}_t\|_\infty \to 0$ almost surely. Consider an estimator of the form $T_n \coloneqq u(\mathrm{P}_n)$, where u is a real-valued statistical functional.

Consider the influence function for u at P_t: Under standard regularity conditions ensuring asymptotic linearity we have

$$u(\mathrm{P}_n) = u(\mathrm{P}_t) + \int \mathrm{IF}(x; u, \mathrm{P}_t)\, (\mathrm{P}_n - \mathrm{P}_t)(\mathrm{d}x) + o_p\Big(n^{-\frac{1}{2}}\Big)\,, \tag{C.7}$$

see also Hampel et al. (1986).

Since $\int \mathrm{IF}(x; u, \mathrm{P}_t)\, \mathrm{dP}_t(x_k) = 0$, this simplifies to

$$u(\mathrm{P}_n) - u(\mathrm{P}_t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}(X_i; u, \mathrm{P}_t) + o_p\Big(n^{-\frac{1}{2}}\Big)\,.$$

After multiplying by $\sqrt{n}$,

$$\sqrt{n}\big(u(\mathrm{P}_n) - u(\mathrm{P}_t)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(X_i; u, \mathrm{P}_t) + o_p(1). \tag{C.8}$$

By the central limit theorem,

$$\sqrt{n}\big(u(\mathrm{P}_n) - u(\mathrm{P}_t)\big) \xrightarrow{d} \mathcal{N}\left(0,\ V(u, \mathrm{P}_t)\right), \qquad V(u, \mathrm{P}_t) \coloneqq \int \mathrm{IF}(x; u, \mathrm{P}_t)^2\, \mathrm{dP}_t(x).$$

Suppose now that $\{\mathrm{P}_\lambda : \lambda \in \Lambda\}$ is a correctly specified regular parametric model, differentiable in quadratic mean, and that u is Fisher consistent, $u(\mathrm{P}_\lambda) = \lambda$. The Fisher information is given by

$$J(\mathrm{P}_\lambda) = \int \left(\frac{\partial}{\partial \lambda} \log p_\lambda(x) \right)^2 \mathrm{dP}_\lambda(x).$$

Differentiating the identity $\int \mathrm{IF}(x; u, \mathrm{P}_\lambda)\, d\mathrm{P}_\lambda(x) = 0$ with respect to λ and using Fisher consistency yields

$$1 = \int \mathrm{IF}(x; u, \mathrm{P}_\lambda) \frac{\partial}{\partial \lambda} \log p_\lambda(x_k)\, d\mathrm{P}_\lambda(x_k). \tag{C.9}$$

Applying the Cauchy–Schwarz inequality to (C.9) gives

$$V(u, \mathrm{P}_\lambda) = \int \mathrm{IF}(x; u, \mathrm{P}_\lambda)^2\, d\mathrm{P}_\lambda(x_k) \;\geq\; \frac{1}{J(\mathrm{P}_\lambda)},$$

which is the Cramér–Rao lower bound. Equality holds if and only if

$$\mathrm{IF}(x; u, \mathrm{P}_\lambda) \propto \frac{\partial}{\partial \lambda} \log p_\lambda(x_k) \quad \mathrm{P}_\lambda\text{-a.s.}$$

For the Barron location functional, the influence function is proportional to the score ψ_k^{B}. Hence, boundedness of ψ_k^{B} (which holds for $\gamma \leq 1$) implies bounded influence and B-robustness in the sense of Hampel et al. (1986). In a Gaussian model, only $\gamma = 2$ yields proportionality to the likelihood score, whereas $\gamma < 2$ induces a deviation from the score and therefore a loss of efficiency.

C.7 Robustness comparison between Calvet et al. (2015) and BLADE

Calvet et al. (2015), evaluate the robustness of a filter in a state space model by requiring that the effect of contamination on the filtered state is uniformly bounded by a linear function of the perturbation size. Their robustness criterion is a pointwise worst-case bound

on the filtering error at the realized contaminated observation, with no separate evaluation draw. By contrast, BLADE ranks two updates by their expected divergence reduction, where the updating draw and the evaluation draw are independent.

Calvet et al. (2015) model contamination as additive in the observation: $y_t = y_t^* + \eta\, u_t$, where y_t^* is drawn from the clean density and η is a perturbation magnitude, and every observation is distorted. The BLADE framework instead uses a mixture $\mathrm{P}_t^{\varepsilon} = (1-\varepsilon)\mathrm{P}_t + \varepsilon \mathrm{H}_t$, where ε is a contamination *fraction*: a proportion ε of draws comes from an arbitrary contaminating measure H_t and the remaining fraction $1-\varepsilon$ is non-contaminated .

A filter in Calvet et al. (2015) is called *robust* if there exists $C_1 \in \mathbb{R}_+$ such that

$$\left|\log \lambda(x_t|y_t, Y_{t-1}) - \log \lambda_{\mathrm{cont}}(x_t|y_t, Y_{t-1}, \eta)\right| \leq C_1|\eta|$$

for all $x_t \in \mathcal{X}$, $y_t \in \mathbb{R}^p$, $Y_{t-1} \in \mathbb{R}^{(t-1)p}$, and $\eta \in I$. This is a worst-case property of the filtering map, bounding how badly any single contaminated observation can distort the posterior. The BLADE robustness result in Theorem 2 instead establishes an *expected* comparison between two updates, if $\varepsilon \geq \underline{\varepsilon}(\gamma_1, \gamma_2, \alpha)$ then $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_{\gamma_1}^{\mathrm{B}}) \geq \Delta\Delta\mathbb{D}_S^{\mathrm{P}_t^{\varepsilon}}(\phi_{\gamma_2}^{\mathrm{B}})$, ranking two updates relative to each other in terms of their expected local S-divergence reduction. Whereas Calvet et al. (2015) view robustness as a stability property of the filtering map under contamination, Theorem 2 views robustness as a comparative property: which of two updates reduces the expected divergence more when the updating draw is contaminated?

C.8 QLE volatility model

The Monte Carlo study on a volatility model was estimated by QLE as described in Section 4. For this the following results are needed to compute $\frac{\partial \hat{\vartheta}_{t+1}(\lambda)}{\partial \lambda}$ where $\lambda = (\omega, \alpha, \beta, \gamma, \xi)$.

The derivative of $\hat{\vartheta}_{t|t-1}(\lambda)$ w.r.t. λ is given by

$$\frac{\partial \hat{\vartheta}_{t+1}(\lambda)}{\partial \lambda} = \frac{\partial \omega}{\partial \lambda} + \psi_k^{\mathrm{B}} \frac{\partial \alpha}{\partial \lambda} + \alpha \frac{\partial \psi_k^{\mathrm{B}}}{\partial \lambda} + \hat{\vartheta}_{t|t-1}(\lambda) \frac{\partial \beta}{\partial \lambda} + \left(\alpha \frac{\partial \psi_k^{\mathrm{B}}}{\partial \vartheta} + \beta \right) \frac{\partial \hat{\vartheta}_{t|t-1}(\lambda)}{\partial \lambda}.$$

Here $\frac{\partial \omega}{\partial \lambda} = (1,0,0,0,0)'$, $\psi_k^{\mathrm{B}} \frac{\partial \alpha}{\partial \lambda} = (0, \psi_k^{\mathrm{B}}, 0, 0, 0)'$, $\alpha \frac{\partial \psi_k^{\mathrm{B}}}{\partial \lambda} = (0,0,0, \alpha \frac{\partial \psi}{\partial \gamma}, \alpha \frac{\partial \psi}{\partial \xi})'$, $\hat{\vartheta}_{t|t-1}(\lambda) \frac{\partial \beta}{\partial \lambda} = (0,0, \hat{\vartheta}_{t|t-1}, 0, 0)'$. When we set $x_k = x_k(y_t, \vartheta_{t|t-1}) = y_t^2 - \vartheta_{t|t-1}$, the second order derivative needed are

$$\frac{\partial \tilde{\psi}_{\gamma,\xi}^{\mathrm{B}}}{\partial \vartheta} = \frac{\partial^2 \tilde{S}_{\gamma,\xi}^{\mathrm{B}}}{\partial x_k^2} = \begin{cases} \dfrac{1}{\xi^2}, & \gamma = 2, \\ \dfrac{1}{\xi^2} \dfrac{1 - \frac{1}{2}(x_k/\xi)^2}{\left(1 + \frac{1}{2}(x_k/\xi)^2\right)^2}, & \gamma = 0, \\ \dfrac{1}{\xi^2} \left(1 - \frac{x_k^2}{\xi^2}\right) \exp\left(-\frac{1}{2}(x_k/\xi)^2\right), & \gamma = -\infty, \\ \dfrac{1}{\xi^2} \left(\frac{(x_k/\xi)^2}{|\gamma-2|} + 1 \right)^{\frac{\gamma}{2}-2} \left[\frac{(x_k/\xi)^2}{|\gamma-2|} + 1 \ + \ \left(\frac{\gamma}{2} - 1\right) \frac{2e^2}{\xi^2 |\gamma-2|} \right], & \text{otherwise.} \end{cases}$$

$$\frac{\partial \tilde{\psi}_k^{\mathrm{B}}}{\partial \gamma} = \frac{x_k \left(\frac{x_k^2}{\xi^2|\gamma-2|} + 1 \right)^{\frac{\gamma}{2}-1} \left(\frac{\ln\left(\frac{x_k^2}{\xi^2|\gamma-2|} + 1 \right)}{2} - \frac{x_k^2 \left(\frac{\gamma}{2} - 1 \right)}{\xi^2 \left(\frac{x_k^2}{\xi^2|\gamma-2|} + 1 \right) |\gamma-2|(\gamma-2)} \right)}{\xi^2} \quad , \gamma \neq 0, 2, -\infty$$

$$\frac{\partial \tilde{\psi}_k^{\mathrm{B}}}{\partial \xi} = \begin{cases} -\dfrac{2e}{\xi^3}, & \gamma = 2, \\ -\dfrac{8 x_k \xi}{\left(x_k^2 + 2c^2\right)^2}, & \gamma = 0, \\ \left[-\dfrac{2e}{\xi^3} + \dfrac{x_k^3}{\xi^5} \right] \exp\left(-\frac{1}{2} \left(\frac{x_k}{\xi} \right)^2 \right), & \gamma = -\infty, \\ \left[-\dfrac{2e}{\xi^3} \ + \ \dfrac{x_k}{\xi^2} \left(\frac{\gamma}{2} - 1 \right) \left(\frac{x_k^2}{\xi^2|\gamma-2|} + 1 \right)^{\frac{\gamma}{2}-2} \cdot (-2) \frac{x_k^2}{\xi^3|\gamma-2|} \right] \left(\frac{x_k^2}{\xi^2|\gamma-2|} + 1 \right)^{\frac{\gamma}{2}-1}, & \text{otherwise.} \end{cases}$$

D Proofs of the results in Appendix A

This section verifies Lemmas 1-4 in Blasques et al. (2023), this done by the verification of Lemma A.1–A.4.

D.1 Verification of Lemma A.1, stationarity and ergodicity

Proof. We separately check condition (i) and (ii) of Lemma A.1.

Proof of Condition (i)

$$\mathbb{E}_{\bar{\mathrm{P}}}\Big[\log^{+}\big|\psi_k^{\mathrm{B}}\big(g(\vartheta^0,\varepsilon_t),\vartheta^0;\gamma,\xi\big)\big|\Big]<\infty,\quad \vartheta^0\in\Theta\subset\mathbb{R};$$

For the proposed model, we have that $y_t = g(\vartheta_{t|t-1},\varepsilon_t)$ and the update equation of $\vartheta_{t+1|t} = \varphi(z_t,\vartheta_{t|t-1}) = \omega+\alpha\psi_k^{\mathrm{B}}(y_t,\vartheta_{t|t-1})+\beta\vartheta_{t|t-1}$. Assume $x_k(y,\vartheta)$ to be finite. Furthermore in this proposition ϑ^0 is to be held constant. ψ_k^{B} is given by

$$\psi_k^{\mathrm{B}}(y,\vartheta^0;\gamma,\xi) = -\frac{\partial\tilde{S}^{\mathrm{B}}_{\gamma,\xi}\big(x_k(y,\vartheta^0)\big)}{\partial x_k(y,\vartheta)}\,\frac{\partial x_k(y,\vartheta^0)}{\partial\vartheta},$$

so

$$\begin{aligned}|\psi_k^{\mathrm{B}}(y,\vartheta^0;\gamma,\xi)| &= \left|-\frac{\partial\tilde{S}^{\mathrm{B}}_{\gamma,\xi}\big(x_k(y,\vartheta^0)\big)}{\partial x_k(y,\vartheta)}\,\frac{\partial x_k(y,\vartheta^0)}{\partial\vartheta}\right|\\ &= \left|\frac{\partial\tilde{S}^{\mathrm{B}}_{\gamma,\xi}\big(x_k(y,\vartheta^0)\big)}{\partial x_k(y,\vartheta)}\,\frac{\partial x_k(y,\vartheta^0)}{\partial\vartheta}\right|\\ &= \left|\frac{\partial\tilde{S}^{\mathrm{B}}_{\gamma,\xi}\big(x_k(y,\vartheta^0)\big)}{\partial x_k(y,\vartheta)}\right|\cdot\left|\frac{\partial x_k(y,\vartheta^0)}{\partial\vartheta}\right|.\end{aligned}$$

In this paper $x_k(y^k,\vartheta) = y^k-\vartheta$. Hence, $\left|\frac{\partial x_k(y,\vartheta^0)}{\partial\vartheta}\right| = 1$ which means that the score reduces to

$$|\psi_k^{\mathrm{B}}(y,\vartheta^0;\gamma,\xi)| = \left|\frac{\partial\tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta^0))}{\partial x_k}\right|.$$

For $\gamma \leq 1$, the Barron score is uniformly bounded in x_k, implying that $|\psi_k^{\mathrm{B}}(y, \vartheta^0; \gamma, \xi)| \leq C$ for some constant $C > 0$. Consequently,

$$\log^+ |\psi_k^{\mathrm{B}}(y, \vartheta^0; \gamma, \xi)| \leq \log C < \infty,$$

and hence

$$\mathbb{E}_{\bar{\mathrm{P}}}\big[\log^+ |\psi_k^{\mathrm{B}}(Y, \vartheta^0; \gamma, \xi)|\big] < \infty.$$

For $1 < \gamma \leq 2$, the Barron score grows at most linearly in x_k, so that $\left|\tilde{\psi}_{\gamma,\xi}^{\mathrm{B}}(x_k)\right| \leq \frac{|x_k|}{\xi^2}$.

This can be written as

$$|\psi_k^{\mathrm{B}}| \leq \frac{|Y_t^k - \vartheta^0|}{\xi^2}.$$

Taking logs we obtain

$$\begin{aligned}
\log^+ |\psi_k^{\mathrm{B}}| &\leq \log^+ \left(\frac{|Y_t^k - \vartheta^0|}{\xi^2}\right) \\
&= \log^+ \xi^{-2} + \log^+ |Y_t^k - \vartheta^0| \\
&\leq \log^+ \xi^{-2} + \log^+ |Y_t^k| + \log^+ |\vartheta^0| \\
&\leq \log^+ \xi^{-2} + k \log^+ |Y_t| + \log^+ |\vartheta^0|.
\end{aligned}$$

Hence for Condition (i) to hold it is sufficient that the following moment condition hold, $\mathbb{E}_{\bar{\mathrm{P}}} \log^+ |Y_t| < \infty$.

Proof of Condition (ii)

To verify:

$$\mathbb{E}_{\bar{\mathrm{P}}}\big[\log \mathfrak{L}_t\big] < 0, \quad \mathfrak{L}_t = \sup_{\vartheta \in \Theta} \Big| \alpha \, \frac{\partial}{\partial \vartheta} \psi_k^{\mathrm{B}}\big(g(\vartheta, \varepsilon_t), \vartheta; \gamma, \xi\big) + \beta \Big|. \tag{D.1}$$

First, consider the second-order derivative

$$\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right| = \left|\frac{\partial^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))}{\partial x_k(y,\vartheta)^2}\cdot\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{\partial \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))}{\partial x_k(y,\vartheta)}\cdot\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right|.$$

The following expression can be simplified

$$\sup_{\vartheta}\left|\alpha\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}+\beta\right|,$$

where ϑ^* is the ϑ that maximizes the expression. By the triangle inequality, we have that

$$\sup_{\vartheta}\left|\alpha\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}+\beta\right| \leq |\alpha|\ \sup_{\vartheta}\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right| + |\beta|\,.$$

For stationarity to hold and by Jensen's inequality (because $\log(\cdot)$ is concave),

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\log\left(|\alpha|\ \sup_{\vartheta}\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right| + |\beta|\right)\right] \leq \log\left(\mathbb{E}_{\bar{\mathrm{P}}}\left[|\alpha|\ \sup_{\vartheta}\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right| + |\beta|\right]\right) < 0,$$

which is equivalent to

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[|\alpha|\ \sup_{\vartheta}\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right| + |\beta|\right] = |\alpha|\ \sup_{\vartheta}\ \mathbb{E}_{\bar{\mathrm{P}}}\left[\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right|\right] + |\beta| < 1.$$

Barron (2019) derives the following conditions

$$\left|\frac{\partial^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))}{\partial x_k(y,\vartheta)^2}\right| \leq \frac{1}{\xi^2},$$

$$\left|\frac{\partial \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta)}{\partial x_k(y,\vartheta)}\right| \leq \begin{cases} \frac{1}{\xi} & \text{if } \gamma \leq 1, \\ \frac{|x_k(y,\vartheta)|}{\xi^2} & \text{if } \gamma \leq 2, \end{cases} \tag{D.2}$$

which results in the following bounds:

$$\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right| \leq \frac{1}{\xi^2}\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{|x_k(y,\vartheta)|}{\xi^2}\left|\left(\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right)\right| \quad \text{if } \gamma \leq 2,$$

and

$$\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right| \leq \frac{1}{\xi^2}\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{1}{\xi}\left|\left(\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right)\right| \quad \text{if } \gamma \leq 1.$$

Case 1: $\gamma \leq 1$

Using the bound for $\gamma \leq 1$ gives

$$\begin{aligned}|\alpha| \sup_{\vartheta\in\Theta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right|\right] &\leq |\alpha| \sup_{\vartheta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\frac{1}{\xi^2}\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{1}{\xi}\left|\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right|\right] \\ &= \frac{|\alpha|}{\xi^2} \sup_{\vartheta\in\Theta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi}\mathbb{E}_{\bar{\mathrm{P}}}\left[\left|\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right|\right].\end{aligned}$$

For stationarity, we need

$$\frac{|\alpha|}{\xi^2} \sup_{\vartheta\in\Theta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi}\mathbb{E}_{\bar{\mathrm{P}}}\left[\left|\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right|\right] + |\beta| < 1.$$

Case 2: $1 < \gamma \leq 2$

Using the bound for $\gamma \leq 2$ one finds

$$\begin{aligned}|\alpha| \sup_{\vartheta\in\Theta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\left|\frac{\partial \psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial \vartheta}\right|\right] &\leq |\alpha| \sup_{\vartheta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\frac{1}{\xi^2}\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{|x_k(y,\vartheta)|}{\xi^2}\left|\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right|\right] \\ &= \frac{|\alpha|}{\xi^2} \sup_{\vartheta\in\Theta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\left(\frac{\partial x_k(y,\vartheta)}{\partial \vartheta}\right)^2\right] \\ &\quad + \frac{|\alpha|}{\xi^2}\mathbb{E}_{\bar{\mathrm{P}}}\left[|x_k(y,\vartheta)| \cdot \left|\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2}\right|\right].\end{aligned}$$

For stationarity, we need

$$\frac{|\alpha|}{\xi^2} \sup_{\vartheta\in\Theta} \mathbb{E}_{\bar{\mathrm{P}}}\left[\left(\frac{\partial x_k(y,\vartheta)}{\partial\vartheta}\right)^2\right] + \frac{|\alpha|}{\xi^2}\mathbb{E}_{\bar{\mathrm{P}}}\left[|x_k(y,\vartheta)|\cdot\left|\frac{\partial^2 x_k(y,\vartheta)}{\partial\vartheta^2}\right|\right] + |\beta| < 1.$$

Using that $x_k(y,\vartheta) \coloneqq y_t^k - \vartheta_{t|t-1}$, we can simplify the conditions using,

$$\frac{\partial x_k(y,\vartheta)}{\partial\vartheta} = -1, \quad \frac{\partial^2 x_k(y,\vartheta)}{\partial\vartheta^2} = 0.$$

This simplifies our conditions to: for both $\gamma \leq 1$ and $1 < \gamma \leq 2$, $\frac{|\alpha|}{\xi^2} + |\beta| < 1$. □

D.2 Verification of Lemma A.2, existence of bounded unconditional moments

Proof. **Proof of Condition (i)**

If one assumes that ε_t is i.i.d, $\{\mathfrak{L}_t\}$ is i.i.d., we verify that

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\left|\psi_k^{\mathrm{B}}\big(g(\vartheta^0,\varepsilon_t),\vartheta^0;\gamma,\xi\big)\right|^r\right] < \infty$$

holds for general $r > 0$. Recall that $x_k(y,\vartheta) \coloneqq y^k - \vartheta$, so $|\partial_\vartheta x_k| = 1$. For $\gamma \leq 1$, the Barron score derivative is uniformly bounded, $|\partial_{x_k}\tilde{S}^{\mathrm{B}}_{\gamma,\xi}| \leq 1/\xi$, so $|\psi_k^{\mathrm{B}}|^r \leq \xi^{-r} < \infty$ for any $r > 0$, and no moment condition on Y_t is required.

For $1 < \gamma \leq 2$, the score derivative grows at most linearly, $|\partial_{x_k}\tilde{S}^{\mathrm{B}}_{\gamma,\xi}| \leq |x_k|/\xi^2$, giving $|\psi_k^{\mathrm{B}}|^r \leq |x_k(y,\vartheta^0)|^r/\xi^{2r}$. Taking expectations, condition (i) holds provided that

$$\mathbb{E}_{\bar{\mathrm{P}}}\Big[\big|Y_t\big|^{kr}\Big] < \infty.$$

Proof of Condition (ii)

We need to show that

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\sup_{\vartheta\in\Theta}\left|\frac{\partial\psi_k^{\mathrm{B}}\big(g(\vartheta,\varepsilon_t),\vartheta;\gamma,\xi\big)}{\partial\vartheta}\right|^r\right]<\infty.$$

The derivative of the score expands as

$$\frac{\partial\psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial\vartheta}=\frac{\partial^2\tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))}{\partial x_k^2}\left(\frac{\partial x_k(y,\vartheta)}{\partial\vartheta}\right)^2+\frac{\partial\tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))}{\partial x_k}\cdot\frac{\partial^2 x_k(y,\vartheta)}{\partial\vartheta^2}.$$

Using $x_k(y,\vartheta)=y^k-\vartheta$, we have $\partial_\vartheta x_k=-1$ and $\partial_\vartheta^2 x_k=0$, so this simplifies to

$$\frac{\partial\psi_k^{\mathrm{B}}(y,\vartheta;\gamma,\xi)}{\partial\vartheta}=\frac{\partial^2\tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k(y,\vartheta))}{\partial x_k^2}.$$

Since $|\partial^2_{x_k}\tilde{S}^{\mathrm{B}}_{\gamma,\xi}|\le 1/\xi^2$ (Barron 2019), the derivative is uniformly bounded in both y and ϑ, giving

$$\sup_{\vartheta\in\Theta}\left|\frac{\partial\psi_k^{\mathrm{B}}}{\partial\vartheta}\right|^r\le\frac{1}{\xi^{2r}}<\infty\quad\text{for any } r>0.$$

Taking expectations immediately yields condition (ii).

Since both conditions (i) and (ii) hold for the same $r>0$, the general result of Blasques et al. (2023) guarantees that $\mathbb{E}_{\bar{\mathrm{P}}}[|\vartheta_{t|t-1}|^s]<\infty$ for some $s>0$, establishing the existence of bounded unconditional moments. □

D.3 Verification of Lemma A.3, invertibility of the BLADE filter

Proof. **Proof of Condition (i)**

For $\gamma\le 1$, the Barron score is uniformly bounded in x_k, so

$$\mathbb{E}_{\bar{\mathrm{P}}}\big[\log^+|\psi_k^{\mathrm{B}}(Y_t,\vartheta^0;\gamma,\xi)|\big]<\infty$$

holds without further assumptions.

For $1 < \gamma \leq 2$, the score grows at most linearly in x_k, i.e.

$$|\psi_k^{\mathrm{B}}(Y_t, \vartheta^0; \gamma, \xi)| \leq \frac{|Y_t^k - \vartheta^0|}{\xi^2}.$$

Hence

$$\log^+ |\psi_k^{\mathrm{B}}| \leq \log^+ \xi^{-2} + \log^+ |Y_t^k - \vartheta^0| \leq \log^+ \xi^{-2} + k \log^+ |Y_t| + \log^+ |\vartheta^0|.$$

Since ϑ^0 is fixed, it follows that

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+ |\psi_k^{\mathrm{B}}|\right] < \infty$$

whenever $\mathbb{E}_{\bar{\mathrm{P}}} \log^+ |Y_t| < \infty$.

Proof of Condition (ii)

Condition (ii) is very similar to the verification of Lemma A.1. By the triangle inequality, we then obtain

$$\sup_{\vartheta \in \Theta} \sup_{\lambda \in \Lambda} \left| \alpha \frac{\partial \psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)}{\partial \vartheta} + \beta \right| \leq |\alpha| \sup_{\vartheta \in \Theta} \sup_{\lambda \in \Lambda} \left| \frac{\partial \psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)}{\partial \vartheta} \right| + |\beta| \,.$$

A sufficient Condition (ii) to be satisfied is when $|\alpha| \sup_{\vartheta \in \Theta} \sup_{\lambda \in \Lambda} \left| \frac{\partial \psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)}{\partial \vartheta} \right| + |\beta| < 1$.

Using the bounds on the derivatives of the Barron loss function, and noting that $\frac{\partial x_k(y,\vartheta)}{\partial \vartheta} = -1$ and $\frac{\partial^2 x_k(y,\vartheta)}{\partial \vartheta^2} = 0$, we obtain

$$\sup_{\vartheta \in \Theta} \sup_{\lambda \in \Lambda} \left| \frac{\partial \psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)}{\partial \vartheta} \right| \leq \sup_{x_k \in \mathbb{R}} \left| \frac{\partial^2 \tilde{S}_{\gamma,\xi}^{\mathrm{B}}(x_k)}{\partial x_k^2} \right| \leq \frac{1}{\xi^2}.$$

Consequently,

$$|\alpha| \sup_{\vartheta \in \Theta} \sup_{\lambda \in \Lambda} \left| \frac{\partial \psi_k^{\mathrm{B}}(y, \vartheta; \gamma, \xi)}{\partial \vartheta} \right| + |\beta| \leq \frac{|\alpha|}{\xi^2} + |\beta|.$$

Therefore, the sufficient condition

$$\frac{|\alpha|}{\xi^2} + |\beta| < 1$$

implies that Condition (ii) is satisfied.

We now have confirmed that the DGP is invertible, since the conditions of Lemma 3 from Blasques et al. (2023) are satisfied.

□

D.4 Verification of Lemma A.4, invertibility properties for the derivatives of the filter

For the proofs of the following conditions we use Table D.1 for the explicit forms of the derivatives used.

Proof. **Proof of Condition (i)**

To check condition one, we split the expectation into the four separate parts and evaluate them separately:

$$\begin{aligned}
&\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+|\psi^{\mathrm{B}}_{\gamma,\xi}| + \log^+\left\|\frac{\partial\psi^{\mathrm{B}}_k}{\partial\lambda}\right\| + \log^+\left|\frac{\partial\psi^{\mathrm{B}}_k}{\partial\vartheta}\right| + \log^+|\vartheta_{t|t-1}(\lambda)|\right] \\
&= \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+|\psi^{\mathrm{B}}_k|\right] + \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\left\|\frac{\partial\psi^{\mathrm{B}}_k}{\partial\lambda}\right\|\right] + \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\left|\frac{\partial\psi^{\mathrm{B}}_k}{\partial\vartheta}\right|\right] + \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+|\vartheta_{t|t-1}(\lambda)|\right].
\end{aligned}$$

From left to right, the first term $\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+|\psi^{\mathrm{B}}_k|\right] < \infty$ by the same arguments as in Lemma A.1, sufficient conditions are $\mathbb{E}_{\bar{\mathrm{P}}}\log^+|Y_t| < \infty$ and $\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\sup_{\lambda\in\Lambda}|\vartheta_{t|t-1}(\lambda)|\right] < \infty$. The second term is $\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\left\|\frac{\partial\psi^B_k}{\partial\lambda}\right\|\right]$. Since γ is fixed, only the ξ-component is non-zero. Writing $x_k = x_k(Y_t, \vartheta_{t|t-1})$ and $u = 1 + x_k^2/(|\gamma-2|\xi^2)$, differentiation of $f(x_k,\xi) = x_k\xi^{-2}u^{\gamma/2-1}$ with respect to ξ gives

$$\frac{\partial\psi^B_k}{\partial\xi} \;=\; -\frac{2x_k}{\xi^3}\,u^{\gamma/2-1} \;-\; \frac{2\left(\frac{\gamma}{2}-1\right)x_k^3}{|\gamma-2|\,\xi^5}\,u^{\gamma/2-2}.$$

Both terms have rate $O(|x_k|^{\gamma-1})$ as $|x_k| \to \infty$, so the derivative grows at most polynomially in x_k. Hence there exist constants $C_1, C_2 > 0$ depending only on Λ and $\underline{\xi}$ such that

$$\log^+ \left\| \frac{\partial \psi_k^B}{\partial \xi} \right\| \;\le\; C_1 + C_2 \log^+ |x_k|.$$

Since $x_k = y^k - \vartheta_{t|t-1}$ is linear in Y_t, we have $\log^+ |x_k| \le \log 2 + C_3 + k \log^+ |Y_t| + \log^+ |\vartheta_{t|t-1}|$ for some finite constant C_3, and therefore

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+ \left\| \frac{\partial \psi_k^B}{\partial \xi} \right\| \right] \;\le\; C_1 + C_2(\log 2 + C_3) + kC_2\, \mathbb{E}_{\bar{\mathrm{P}}}[\log^+ |Y_t|] + C_2\, \mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+ \sup_{\lambda \in \Lambda} |\vartheta_{t|t-1}(\lambda)| \right] \;<\; \infty,$$

under the maintained conditions $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+ |Y_t|] < \infty$ and $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+ \sup_\lambda |\vartheta_{t|t-1}(\lambda)|] < \infty$.

Proof of Condition (ii)

Condition (ii) requires

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+ \left(\sup_{\vartheta} \left| \frac{\partial \psi_k^B}{\partial \vartheta} \right| + \sup_{\vartheta, \lambda} \left\| \frac{\partial^2 \psi_k^B}{\partial \lambda\, \partial \vartheta} \right\| + \sup_{\vartheta} \left\| \frac{\partial^2 \psi_k^B}{\partial \vartheta^2} \right\| + \sup_{\lambda} \| \vartheta'_{t|t-1}(\lambda) \| \right) \right] < \infty.$$

Using $\log^+(x_1 + x_2 + x_3 + x_4) \le \log 4 + \sum_{i=1}^4 \log^+ x_i$, it suffices to show each of the four expectations is finite.

First term: $\partial_\vartheta \psi_k^B$. By the chain rule,

$$\frac{\partial \psi_k^B(y, \vartheta; \gamma, \xi)}{\partial \vartheta} = \frac{\partial^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k)}{\partial x_k^2} \left(\frac{\partial x_k}{\partial \vartheta} \right)^2 + \frac{\partial \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k)}{\partial x_k} \frac{\partial^2 x_k}{\partial \vartheta^2}.$$

Since $x_k = y^k - \vartheta$ we have $\partial_\vartheta x_k = -1$ and $\partial_\vartheta^2 x_k = 0$, so this reduces to $\partial_\vartheta \psi_k^B = \partial_{x_k}^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k)$. From the explicit Barron score, $\sup_{x_k \in \mathbb{R}} |\partial_{x_k}^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k)| \le \xi^{-2}$ for all $(\gamma, \xi) \in \Lambda$ with $\gamma \le 2$ and $\xi \ge \underline{\xi} > 0$. Hence

$$\sup_{\vartheta \in \Theta} \left| \frac{\partial \psi_k^B}{\partial \vartheta} \right| \;\le\; \frac{1}{\underline{\xi}^2} \;<\; \infty,$$

and $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+(\xi^{-2})] < \infty$ since ξ is a deterministic constant.

Second term: $\partial_\lambda \partial_\vartheta \psi_k^B$. Using $\partial_\vartheta x_k = -1$ and $\partial_\vartheta^2 x_k = 0$ as above, $\partial_\vartheta \psi_k^B = \partial_{x_k}^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k)$. Differentiating with respect to $\lambda = \xi$, and noting that $x_k = y^k - \vartheta$ does not depend on ξ, the operator ∂_ξ acts only on the explicit ξ-dependence in $\tilde{S}^{\mathrm{B}}_{\gamma,\xi}$:

$$\frac{\partial^2 \psi_k^B}{\partial \xi \, \partial \vartheta} = \frac{\partial}{\partial \xi}\Big(\partial_{x_k}^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k)\Big) = \partial_\xi \partial_{x_k}^2 \tilde{S}^{\mathrm{B}}_{\gamma,\xi}(x_k).$$

This is a finite sum of terms of the form $c_j(\gamma,\xi) \cdot x_k^{n_j} \cdot u^{\gamma/2 - r_j}$ with $u = 1 + x_k^2/(|\gamma - 2|\xi^2)$ and rate exponent $n_j + \gamma - 2r_j = \gamma - 2 \le 0$, see Table D.1, since there is one ϑ-differentiation and $\gamma \le 2$. Hence each term is uniformly bounded on $\mathbb{R}$. Since Λ is compact with $\xi \ge \underline{\xi} > 0$ and $|\gamma - 2| \ge \underline{c} > 0$ for $\gamma < 2$ (when $\gamma = 2$ the derivative is equal to 0, see Remark 1), the prefactors $c_j(\gamma,\xi)$ are bounded on Λ, giving

$$\sup_{\vartheta \in \Theta,\, \lambda \in \Lambda} \left\| \frac{\partial^2 \psi_k^B}{\partial \lambda \, \partial \vartheta} \right\| \;\le\; C \;<\; \infty,$$

and $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+ C] < \infty$ follows immediately.

Third term: $\partial_{\vartheta^2}^2 \psi_k^B$. Using $\partial_\vartheta = -\partial_{x_k}$ on $f(x_k,\xi) = \psi_k^B$,

$$\frac{\partial^2 \psi_k^B}{\partial \vartheta^2} = \frac{\partial^2 f}{\partial x_k^2}(x_k,\xi),$$

which is a finite sum of terms $c_j(\gamma,\xi) \cdot x_k^{n_j} \cdot u^{\gamma/2 - r_j}$ with rate exponent $n_j + \gamma - 2r_j = \gamma - 3 \le -1 < 0$ for $\gamma \le 2$. Explicitly, derived from Table D.1,

$$\frac{\partial^2 \psi_k^B}{\partial x_k^2} = \frac{\operatorname{sgn}(\gamma - 2)\, x_k}{\xi^4}\, u^{\gamma/2 - 3} \left[3 + \frac{(\gamma - 1)\, x_k^2}{|\gamma - 2|\xi^2} \right],$$

which behaves as $O(x_k^{\gamma-3})$ as $|x_k| \to \infty$. Since $\gamma \leq 2$, the exponent $\gamma - 3 \leq -1 < 0$, so

$$\sup_{\vartheta \in \Theta,\, \lambda \in \Lambda} \left\| \frac{\partial^2 \psi_k^B}{\partial \vartheta^2} \right\| \;\leq\; C \;<\; \infty,$$

and $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+ C] < \infty$.

Fourth term: $\sup_\lambda \|\vartheta'_{t|t-1}(\lambda)\|$. We require finiteness of $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+ \sup_\lambda \|\vartheta'_{t|t-1}(\lambda)\|]$ which is guaranteed by Assumption 3

All four terms are finite, so Condition (ii) holds.

Proof of Condition (iii)

Condition (iii) requires, for all $\lambda \in \Lambda$,

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+ \left\| \frac{\partial^2 \psi_k^B}{\partial \lambda \partial \lambda^\top} \right\| + \log^+ \left\| \frac{\partial^2 \psi_k^B}{\partial \lambda \partial \vartheta} \right\| + \log^+ \left\| \frac{\partial^2 \psi_k^B}{\partial \vartheta^2} \right\| \right] < \infty.$$

Since γ is fixed, $\lambda = \xi$ and the three required derivatives are $\partial^2_{\xi^2} \psi_k^B$, $\partial_\xi \partial_\vartheta \psi_k^B$, and $\partial^2_{\vartheta^2} \psi_k^B$. Write $\psi_k^B(y, \vartheta; \gamma, \xi) = f(x_k, \xi)$ where $x_k = y^k - \vartheta$ and $u := 1 + x_k^2/(|\gamma - 2|\xi^2)$.

Terms 2 and 3: $\partial_\xi \partial_\vartheta \psi_k^B$ and $\partial^2_{\vartheta^2} \psi_k^B$. Both derivatives are finite sums of terms of the form $c_j(\gamma, \xi) \cdot x_k^{n_j} \cdot u^{\gamma/2 - r_j}$ with rate exponent $n_j + \gamma - 2r_j = \gamma - 1 - m$, where $m \in \{1, 2\}$ is the number of ϑ-differentiations, see Table D.1. Since $m \geq 1$ and $\gamma \leq 2$, this exponent satisfies $\gamma - 1 - m \leq 0$. When the exponent is strictly negative, $|x_k|^{n_j + \gamma - 2r_j} \to 0$ as $|x_k| \to \infty$, so each term is bounded on $\mathbb{R}$ by continuity; when the exponent equals zero, the term is constant in x_k. In either case, every term is uniformly bounded on $\mathbb{R}$. Since Λ is compact with $\xi \geq \underline{\xi} > 0$ and $|\gamma - 2| \geq \underline{c} > 0$ (the quadratic case $\gamma = 2$ is excluded from the general analysis, with $\xi = 1$ fixed by Remark 1), the prefactors $c_j(\gamma, \xi)$ are bounded on Λ. Hence both suprema are bounded by deterministic constants, and $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+(\cdots)] < \infty$ follows immediately. Term 1: $\partial^2_{\xi^2} \psi_k^B$. The derivative is given by the following expression,

$$\frac{6\left(y^k-\vartheta\right)\left(\frac{\left(y^k-\vartheta\right)^2}{|\gamma-2|\xi^2}+1\right)^{\frac{\gamma}{2}-1}}{\xi^4}+\frac{14\left(\frac{\gamma}{2}-1\right)\left(y^k-\vartheta\right)^3\left(\frac{\left(y^k-\vartheta\right)^2}{|\gamma-2|\xi^2}+1\right)^{\frac{\gamma}{2}-2}}{|\gamma-2|\,\xi^6}$$
$$+\frac{4\left(\frac{\gamma}{2}-2\right)\left(\frac{\gamma}{2}-1\right)\left(y^k-\vartheta\right)^5\left(\frac{\left(y^k-\vartheta\right)^2}{|\gamma-2|\xi^2}+1\right)^{\frac{\gamma}{2}-3}}{(\gamma-2)^2\,\xi^8}$$

Since x_k does not depend on ξ, the operator ∂_ξ acts only on the explicit ξ-dependence in f. The derivative $\partial^2_{\xi^2} f(x_k,\xi)$ is a finite sum of terms $c_j(\gamma,\xi)\cdot x_k^{n_j}\cdot u^{\gamma/2-r_j}$ with $m=0$ ϑ-differentiations. Inspecting the explicit expression, the leading term as $|x_k|\to\infty$ grows like $|x_k|^{\gamma-1}$. Since each term in $\partial^2_{\xi^2}\psi_k^B$ is a finite product of powers of x_k and u, the derivative grows at most polynomially in x_k, so there exist constants $C_1, C_2 > 0$ depending only on Λ and $\underline{\xi}$ such that

$$\log^+\left|\frac{\partial^2\psi_k^B}{\partial\xi^2}\right| \;\le\; C_1 + C_2\log^+|x_k|.$$

Since $x_k = y^k - \vartheta_{t|t-1}$, we have

$$\log^+|x_k| \;\le\; \log 2 + k\log^+|Y_t| + \log^+|\vartheta_{t|t-1}|,$$

and therefore

$$\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\left\|\frac{\partial^2\psi_k^B}{\partial\xi^2}\right\|\right] \;\le\; C_1 + C_2\log 2 + kC_2\,\mathbb{E}_{\bar{\mathrm{P}}}[\log^+|Y_t|] + C_2\,\mathbb{E}_{\bar{\mathrm{P}}}\left[\log^+\sup_{\lambda\in\Lambda}|\vartheta_{t|t-1}(\lambda)|\right] \;<\; \infty,$$

where both expectations are finite by the maintained conditions $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+|Y_t|]<\infty$ and $\mathbb{E}_{\bar{\mathrm{P}}}[\log^+\sup_\lambda|\vartheta_{t|t-1}(\lambda)|]<\infty$ guaranteed in Assumption 3. All three terms are finite, so Condition (iii) holds.

Proof of Condition (iv)

Condition (iv) requires

$$\mathbb{E}_{\bar{\mathrm{P}}_t}\left[\log^+\left(\sup_{\vartheta,\lambda_i,\lambda_j}\left|\frac{\partial^3\psi_t^{\mathrm{B}}}{\partial\lambda_i\partial\lambda_j\partial\vartheta}\right|+\sup_{\vartheta,\lambda_i}\left\|\frac{\partial^3\psi_t^{\mathrm{B}}}{\partial\lambda_i\partial\vartheta^2}\right\|+\sup_{\vartheta}\left\|\frac{\partial^3\psi_t^{\mathrm{B}}}{\partial\vartheta^3}\right\|\right)\right]<\infty. \tag{D.3}$$

Since γ is fixed, $\lambda=\xi$ is the only free parameter, so $\lambda_i=\lambda_j=\xi$ and the three required derivatives are $\partial_{\xi^2}\partial_\vartheta\psi_t^{\mathrm{B}}$, $\partial_\xi\partial_{\vartheta^2}\psi_t^{\mathrm{B}}$, and $\partial_{\vartheta^3}\psi_t^{\mathrm{B}}$.

Write $\psi_k^B(y,\vartheta;\gamma,\xi)=f(x_k,\xi)$ where $x_k=y^k-\vartheta$ and

$$f(x,\xi)\;:=\;\frac{x}{\xi^2}\left(1+\frac{x^2}{|\gamma-2|\xi^2}\right)^{\gamma/2-1}.$$

Since x_k depends on ϑ but not on ξ, the operator ∂_ϑ acts as $-\partial_x$ on $f(x_k,\xi)$, while ∂_ξ acts on the explicit ξ-dependence in f. These two operations commute. Therefore

$$\frac{\partial^3\psi_t^{\mathrm{B}}}{\partial\xi^2\partial\vartheta}=-\frac{\partial^3 f}{\partial\xi^2\partial x}(x_k,\xi),\qquad\frac{\partial^3\psi_t^{\mathrm{B}}}{\partial\xi\,\partial\vartheta^2}=\frac{\partial^3 f}{\partial\xi\,\partial x^2}(x_k,\xi),\qquad\frac{\partial^3\psi_t^{\mathrm{B}}}{\partial\vartheta^3}=-\frac{\partial^3 f}{\partial x^3}(x_k,\xi). \tag{D.4}$$

Setting $u:=1+x_k^2/(|\gamma-2|\xi^2)$, each of the three derivatives in (D.4) is a finite sum of terms of the form

$$T_j(x_k,\xi)\;:=\;c_j(\gamma,\xi)\cdot x_k^{n_j}\cdot u^{\gamma/2-r_j},$$

where $n_j\geq 0$ and $r_j\geq 1$ are non-negative integers and $c_j(\gamma,\xi)$ is a rational function of (γ,ξ) whose explicit form is given in Table D.1. Denoting by m the number of ϑ-differentiations, the index pairs (n_j,r_j) appearing in each derivative are:

Derivative	m	n_j	r_j
$\partial_{\xi^2}\partial_\vartheta \psi_t^{\mathrm{B}}$	1	$\{0,2,4,6\}$	$n_j/2+1$
$\partial_\xi \partial_{\vartheta^2} \psi_t^{\mathrm{B}}$	2	$\{1,3,5\}$	$(n_j+1)/2+1$
$\partial_{\vartheta^3} \psi_t^{\mathrm{B}}$	3	$\{0,2,4\}$	$n_j/2+2$

We show each term T_j is uniformly bounded over $x_k \in \mathbb{R}$. Since $u \geq 1 > 0$ and $\gamma/2 - r_j \leq \gamma/2 - 1 \leq 0$ (as $r_j \geq 1$ and $\gamma \leq 2$), the map $t \mapsto t^{\gamma/2-r_j}$ is non-increasing on $(0,\infty)$. Using $u \geq x_k^2/(|\gamma-2|\xi^2)$ therefore gives

$$u^{\gamma/2-r_j} \;\leq\; \left(\frac{x_k^2}{|\gamma-2|\xi^2}\right)^{\gamma/2-r_j},$$

and hence

$$|T_j(x_k,\xi)| \;\leq\; |c_j(\gamma,\xi)| \cdot (|\gamma-2|\xi^2)^{r_j-\gamma/2} \cdot |x_k|^{n_j+\gamma-2r_j}. \tag{D.5}$$

The exponent of $|x_k|$ satisfies

$$n_j + \gamma - 2r_j \;=\; \gamma - 1 - m \tag{D.6}$$

for every term in the derivative with m ϑ-differentiations. Identity (D.6) holds because each differentiation in x raises n_j by 1 and r_j by 1 simultaneously via the product and chain rules, so $n_j - 2r_j$ decreases by 1 per differentiation, which can be verified directly from Table D.1. Since $m \geq 1$ and $\gamma \leq 2$,

$$n_j + \gamma - 2r_j \;=\; \gamma - 1 - m \;\leq\; \gamma - 2 \;\leq\; 0.$$

When this exponent is strictly negative, $|x_k|^{n_j+\gamma-2r_j} \to 0$ as $|x_k| \to \infty$, so the function

is bounded on $\mathbb{R}$ by continuity. When the exponent equals zero (only at $\gamma = 2$, treated separately as the quadratic case with $\xi = 1$ fixed), the term is constant in x_k. In either case,

$$\sup_{x_k \in \mathbb{R}} |x_k|^{n_j+\gamma-2r_j} \;\le\; K_j \;<\; \infty.$$

The prefactor $(|\gamma-2|\xi^2)^{r_j-\gamma/2}$ in (D.5) is continuous in (γ, ξ) on Λ. Since Λ is compact with $\xi \ge \underline{\xi} > 0$ and $|\gamma - 2| \ge \underline{c} > 0$ (as $\gamma = 2$ is excluded from the general case and handled separately), this prefactor is bounded on Λ. Likewise, $c_j(\gamma, \xi)$ is continuous and hence bounded on Λ. Therefore

$$\sup_{x_k \in \mathbb{R},\, (\gamma,\xi)\in\Lambda} |T_j(x_k, \xi)| \;\le\; \sup_{\Lambda} |c_j| \cdot \sup_{\Lambda} (|\gamma-2|\xi^2)^{r_j-\gamma/2} \cdot K_j \;=:\; C_j \;<\; \infty.$$

Since each of the three derivatives in (D.4) is a finite sum of such terms, the three suprema in (D.3) are each bounded by a deterministic constant $C < \infty$ not depending on Y_t.

It follows that

$$\log^+\left(\sup_{\vartheta,\xi} \left| \frac{\partial^3 \psi_t^{\mathrm{B}}}{\partial \xi^2 \partial \vartheta} \right| + \sup_{\vartheta,\xi} \left| \frac{\partial^3 \psi_t^{\mathrm{B}}}{\partial \xi\, \partial \vartheta^2} \right| + \sup_{\vartheta} \left| \frac{\partial^3 \psi_t^{\mathrm{B}}}{\partial \vartheta^3} \right| \right) \;\le\; \log^+(3C) \;<\; \infty$$

almost surely, and taking $\mathbb{E}_{\bar{\mathrm{P}}}[\,\cdot\,]$ yields (D.3). □

This appendix verified that the BLADE filter satisfies all requirements of Lemmas 1–4 in Blasques et al. (2023). In particular, under mild assumptions, the recursion is shown to admit a stationary and ergodic solution, possess bounded unconditional moments and to be invertible. Moreover, the derivative of the filter are themselves, stationary, ergodic and invertible. These results ensure the existence of well-defined and stable data-generating process for the proposed BLADE filter and justify the use of sample-based approximations. Consequently, the regularity conditions ensure that the results of Theorem 3 and 4 hold. □

Table D.1: Derivatives of $\psi^{\mathrm{B}}_{\gamma,\xi}$.

Derivative	**Full expression**	**Rate**
Pure ϑ-derivatives		
$\dfrac{\partial \psi_k^B}{\partial \vartheta}$	$-\dfrac{(\frac{\gamma}{2}-1)x_k^2}{\frac{\lvert\gamma-2\rvert}{2}\,\xi^4}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-2}-\dfrac{1}{\xi^2}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-1}$	$x_k^{\gamma-2}$
$\dfrac{\partial^2 \psi_k^B}{\partial \vartheta^2}$	$\dfrac{6(\frac{\gamma}{2}-1)x_k}{\lvert\gamma-2\rvert\,\xi^4}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-2}+\dfrac{4(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^3}{(\gamma-2)^2\xi^6}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-3}$	$x_k^{\gamma-3}$
$\dfrac{\partial^3 \psi_k^B}{\partial \vartheta^3}$	$-\dfrac{6(\frac{\gamma}{2}-1)}{\lvert\gamma-2\rvert\,\xi^4}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-2}-\dfrac{24(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^2}{(\gamma-2)^2\xi^6}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-3}-\dfrac{8(\frac{\gamma}{2}-3)(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^4}{\lvert\gamma-2\rvert\,(\gamma-2)^2\xi^8}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-4}$	$x_k^{\gamma-4}$
Mixed ξ-ϑ derivatives		
$\dfrac{\partial^2 \psi_k^B}{\partial \xi\,\partial \vartheta}$	$\dfrac{2}{\xi^3}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-1}+\dfrac{10(\frac{\gamma}{2}-1)x_k^2}{\lvert\gamma-2\rvert\,\xi^5}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-2}+\dfrac{4(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^4}{(\gamma-2)^2\xi^7}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-3}$	$x_k^{\gamma-2}$
$\dfrac{\partial^3 \psi_k^B}{\partial \xi^2\,\partial \vartheta}$	$-\dfrac{6}{\xi^4}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-1}-\dfrac{54(\frac{\gamma}{2}-1)x_k^2}{\lvert\gamma-2\rvert\,\xi^6}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-2}-\dfrac{48(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^4}{(\gamma-2)^2\xi^8}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-3}-\dfrac{8(\frac{\gamma}{2}-3)(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^6}{\lvert\gamma-2\rvert\,(\gamma-2)^2\xi^{10}}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-4}$	$x_k^{\gamma-2}$
$\dfrac{\partial^3 \psi_k^B}{\partial \xi\,\partial \vartheta^2}$	$-\dfrac{24(\frac{\gamma}{2}-1)x_k}{\lvert\gamma-2\rvert\,\xi^5}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-2}-\dfrac{36(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^3}{(\gamma-2)^2\xi^7}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-3}-\dfrac{8(\frac{\gamma}{2}-3)(\frac{\gamma}{2}-2)(\frac{\gamma}{2}-1)x_k^5}{\lvert\gamma-2\rvert\,(\gamma-2)^2\xi^9}\left(\dfrac{x_k^2}{\lvert\gamma-2\rvert\xi^2}+1\right)^{\gamma/2-4}$	$x_k^{\gamma-3}$

NOTE: Full expressions for derivatives of $\psi_k^B(y,\vartheta;\gamma,\xi)$ required for Lemma A.4. Notation: $x_k = y^k - \vartheta$, $u = 1 + x_k^2/(|\gamma-2|\xi^2)$, $\gamma \in (-\infty, 2]$ fixed, $\xi \geq \underline{\xi} > 0$. Each term in every derivative has the form $x_k^n \cdot u^{\gamma/2-r}$. As $|x_k| \to \infty$, $u \sim x_k^2/(|\gamma-2|\xi^2)$, so $x_k^n \cdot u^{\gamma/2-r} \sim C \cdot x_k^{n+\gamma-2r}$, where $C > 0$ depends only on $(\gamma,\xi) \in \Lambda$. Within each derivative, every term shares the same exponent $n+\gamma-2r$, because each ϑ-differentiation raises both n and r by one via the product and chain rules, leaving $n-2r$ unchanged. Consequently, the growth rate depends only on the number of ϑ-derivatives: it equals $x_k^{\gamma-1-m}$ where m is the order of ϑ-differentiation, irrespective of the number of ξ-derivatives. Since $\gamma \leq 2$, all rates are $O(x_k^{-1})$ or faster, hence every derivative is uniformly bounded on $\mathbb{R} \times \Lambda$.

E Evaluation of a scaling matrix using the expected divergence reduction difference

In the classical score drive model literature originally proposed by Creal et al. (2013) and Harvey (2013) (see `https://www.gasmodel.com` for an overview), the use of a scaling matrix S_t in the update equation is common. The update equation pre-multiplies the inner score by a scaling matrix $\mathcal{S}_{t-1}$, measurable with respect to $\mathcal{F}_{t-1}^{Y}$. The canonical family of choices is parametrized by an exponent $\varrho \in \{0,\, 1/2,\, 1\}$ and given by

$$\begin{aligned} \mathcal{S}_{t-1}(\varrho) \;&=\; \left(\int_{\mathcal{Y}} s\big(y, \vartheta_{t|t-1}\big)\, s\big(y, \vartheta_{t|t-1}\big)^{\top} f_{t|t-1}(y)\, \mathrm{d}y \right)^{-\varrho}, \\ s\big(y, \vartheta_{t|t-1}\big) \;&\coloneqq\; \nabla_{\vartheta} \log f\big(y \,\big|\, \vartheta\big)\Big|_{\vartheta=\vartheta_{t|t-1}}, \end{aligned} \tag{E.1}$$

conditional Fisher information $\mathcal{S}_{t-1}(1) = \mathcal{I}_{t-1}^{-1}$ with

$$\mathcal{I}_{t-1} \;\coloneqq\; \mathbb{E}_{\mathrm{F}_{\vartheta_{t|t-1}}}\Big[s\big(Y_t, \vartheta_{t|t-1}\big)\, s\big(Y_t, \vartheta_{t|t-1}\big)^{\top} \Big], \tag{E.2}$$

and $\varrho = 1/2$ yields the inverse symmetric square root $\mathcal{S}_{t-1}(1/2) = \mathcal{I}_{t-1}^{-1/2}$. The integral in (E.1) is taken under the postulated model density $f_{t|t-1}$, so that $\mathcal{S}_{t-1}(\varrho)$ is computable from the model alone and is $\mathcal{F}_{t-1}^{Y}$-measurable by construction. A fourth common choice, outside the family (E.1), is the inverse conditional expected Hessian

$$\mathcal{S}_{t-1}^{\mathrm{H}} \;=\; \mathcal{H}_{t-1}^{-1}, \qquad \mathcal{H}_{t-1} \;\coloneqq\; -\,\mathbb{E}_{\mathrm{F}_{\vartheta_{t|t-1}}}\Big[\nabla_{\vartheta}^{2}\, S\big(F_{\vartheta_{t|t-1}}, Y_t\big) \Big], \tag{E.3}$$

defined relative to the scoring rule S that enters the divergence $\mathbb{D}_S$. The Fisher information (E.2) is a property of the model specification $\mathrm{F}_{\vartheta_{t|t-1}}$ alone, whereas the expected Hessian (E.3) depends on the choice of scoring rule.

The BLADE filter uses the identity scaling $\mathcal{S}_{t-1}(0) = I$ throughout. The choice between

alternative scalings in (E.1)–(E.3) is typically motivated in the score-driven literature by heuristic considerations of efficiency or invariance, rather than by a formal criterion linking the scaling matrix to a divergence-reduction property of the resulting update. In this appendix we observe that the expected divergence reduction difference of Definition 3 supplies such a criterion; holding the inner score fixed and varying only the scaling matrix, the criterion isolates the contribution of $\mathcal{S}_{t-1}$ to the expected divergence reduction.

To formalize this, consider the following setup. Let $\vartheta_{t|t-1} \in \mathbb{R}^k$, extending the time varying parameter to a higher dimensional space. Let $\psi : \mathcal{Y} \times \Theta \to \Theta$ be an inner score and let R_t denote the updating distribution generating the updating draw $\check{Y}_t$. Consider two updates that share ψ and differ only in their scaling matrices $\mathcal{S}_{t-1}$

$$\phi_i\big(y, \vartheta_{t|t-1}\big) = \vartheta_{t|t-1} + A\mathcal{S}_{t-1}^{(i)}\big(\vartheta_{t|t-1}\big)\,\psi\big(y, \vartheta_{t|t-1}\big), \qquad i \in \{1, 2\}, \tag{E.4}$$

with A the learning rate matrix with $A\mathcal{S}_{t-1}^{(i)} \succ O_k$ and $\mathcal{S}_{t-1}^{(1)} \neq \mathcal{S}_{t-1}^{(2)}$ both positive and $\mathcal{F}_{t-1}^{Y}$-measurable. Following Definition 4, we say that ϕ_1 P_t-*contractively dominates* ϕ_2 for the inner score ψ, with respect to $(\mathcal{P}, \mathbb{D}_S, \mathcal{T})$, whenever $\Delta\Delta\mathbb{D}_S^{\mathrm{R}_t}\big(\phi_1 \,\big\|\, \phi_2\big) \geq 0$.

Because ψ is shared between both update rules, the ranking between $\mathcal{S}_{t-1}^{(1)}$ and $\mathcal{S}_{t-1}^{(2)}$ is, at fixed inner score ψ, fixed scoring rule S, and fixed true measure P_t, governed by the pair $\big(\mathcal{S}_{t-1}^{(1)}, \mathcal{S}_{t-1}^{(2)}\big)$ alone. Across configurations, however, ψ, S, and P_t each affect the sign of $\Delta\Delta\mathbb{D}_S^{\mathrm{R}_t}\big(\phi_1 \,\big\|\, \phi_2\big)$; any notion of scaling-matrix optimality is therefore relative to the configuration $\big(\psi, (\mathcal{P}, \mathbb{D}_S, \mathcal{T})\big)$.

E.1 Score-driven updates

For illustrational purposes we specialize to the score-driven updates of Creal et al. (2013) in a univariate setting, for which we obtain the following update structure,

$$\phi_i\big(y, \vartheta_{t|t-1}\big) \;=\; \vartheta_{t|t-1} + \alpha\, \mathcal{S}_{t-1}^{(i)}\, s\big(y, \vartheta_{t|t-1}\big), \qquad i \in \{1, 2\}, \tag{E.5}$$

where $s(y, \vartheta_{t|t-1}) \coloneqq \nabla_\vartheta \log f(y \mid \vartheta)\big|_{\vartheta=\vartheta_{t|t-1}} \in \mathbb{R}$ is the log-score of the postulated model, $\mathcal{S}_{t-1}^{(i)} \in \mathbb{R}$ is $\mathcal{F}_{t-1}^{Y}$-measurable, $\alpha > 0$ is a scalar learning rate, and the scoring rule entering the divergence $\mathbb{D}_S$ is the logarithmic scoring rule, $S = \text{LogS}$. These updates induce the local divergence space $(\mathcal{P}, \text{KL}, \text{id})$, where KL denotes the Kullback–Leibler divergence and id the identity functional.

To interpret the effect of the scaling factor on the update we use the exact Taylor expansion from the proof of Theorem 1. Substituting (E.5) into $\Delta\Delta\mathbb{D}_S^{\mathrm{P}_t}(\phi_1 \,\|\, \phi_2)$ introduced in the proof of Theorem 1 yields

$$\begin{aligned}
\Delta\Delta\mathbb{D}_{\text{LogS}}^{\mathrm{P}_t}(\phi_1 \| \phi_2) &\coloneqq \Delta\mathbb{D}_{\text{LogS}}^{\mathrm{P}_t}(\phi_1) - \Delta\mathbb{D}_{\text{LogS}}^{\mathrm{P}_t}(\phi_2) \\
&= \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t}\Big[\text{LogS}(\mathrm{F}_{\vartheta_{t|t}(\check{Y}_t, \gamma_1)}, Y_t)\Big] - \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t}\Big[\text{LogS}(\mathrm{F}_{\vartheta_{t|t}(\check{Y}_t, \gamma_2)}, Y_t)\Big] \\
&= \; -\alpha\, A_t \;+\; \alpha^2\, Q_t,
\end{aligned} \tag{E.6}$$

with $A_t, Q_t \in \mathbb{R}$ are given by the expressions

$$A_t \;=\; \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t}\Big[\nabla_\vartheta \,\text{LogS}\big(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t\big)\big(\mathcal{S}_{t-1}^{(2)} - \mathcal{S}_{t-1}^{(1)}\big)\, s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\Big], \tag{E.7}$$

$$Q_t \;=\; \mathbb{E}_{\mathrm{P}_t \otimes \mathrm{P}_t}\Big[\widetilde{H}_{\text{LogS}}\big(Y_t, \vartheta_{t|t}^{(1)}\big)\big(\mathcal{S}_{t-1}^{(1)} s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\big)^2 \;-\; \widetilde{H}_{\text{LogS}}\big(Y_t, \vartheta_{t|t}^{(2)}\big)\big(\mathcal{S}_{t-1}^{(2)} s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\big)^2\Big], \tag{E.8}$$

where $\mathcal{S}_{t-1}^{(i)} \coloneqq \mathcal{S}_t(\varrho_i)$, $\vartheta_{t|t}^{(i)} \coloneqq \phi_i(\check{Y}_t, \vartheta_{t|t-1})$, and $\widetilde{H}_{\text{LogS}}(Y_t, \vartheta_{t|t}^{(i)})$ denotes the integrated Hes-

sian

$$\widetilde{H}_{\mathrm{LogS}}\big(Y_t, \vartheta^{(i)}_{t|t}\big) \;\coloneqq\; \int_0^1 (1-u)\, \nabla^2_{\vartheta} \mathrm{LogS}\big(\mathrm{F}_{\vartheta_{t|t-1}+u\alpha\mathcal{S}^{(i)}_{t-1}s\left(\check{Y}_t,\vartheta_{t|t-1}\right)}, Y_t\big)\, du.$$

For the purpose of illustration we compare an update ϕ_1 with $\mathcal{S}^{(1)}_{t-1} = \mathcal{I}^{-1}_{t-1}$ (E.2) and ϕ_2 with $\mathcal{S}^{(2)}_{t-1} = 1$. In this setting (E.6) obtains A_t and Q_t equal to

$$A_t \;=\; \mathbb{E}_{\mathrm{P}_t\otimes\mathrm{P}_t}\Big[\nabla_{\vartheta}\mathrm{LogS}\big(\mathrm{F}_{\vartheta_{t|t-1}}, Y_t\big)\big(1-\mathcal{I}^{-1}_{t-1}\big)\, s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\Big], \tag{E.9}$$

$$Q_t \;=\; \mathbb{E}_{\mathrm{P}_t\otimes\mathrm{P}_t}\Big[\widetilde{H}_{\mathrm{LogS}}\big(Y_t, \vartheta^{(1)}_{t|t}\big)\big(\mathcal{I}^{-1}_{t-1}s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\big)^2 \;-\; \widetilde{H}_{\mathrm{LogS}}\big(Y_t, \vartheta^{(2)}_{t|t}\big)s\big(\check{Y}_t, \vartheta_{t|t-1}\big)^2\Big], \tag{E.10}$$

with updated parameters

$$\vartheta^{(1)}_{t|t} \;=\; \vartheta_{t|t-1} + \alpha\, \mathcal{I}^{-1}_{t-1}\, s\big(\check{Y}_t, \vartheta_{t|t-1}\big), \qquad \vartheta^{(2)}_{t|t} \;=\; \vartheta_{t|t-1} + \alpha\, s\big(\check{Y}_t, \vartheta_{t|t-1}\big).$$

First-order term. Using $\nabla_{\vartheta}\mathrm{LogS}(F_{\vartheta}, y) = s(y, \vartheta)$ for the logarithmic scoring rule, the $\mathcal{F}^{Y}_{t-1}$-measurability of $\mathcal{I}^{-1}_{t-1}$, and independence of Y_t and $\check{Y}_t$ under $\mathrm{P}_t \otimes \mathrm{P}_t$, A_t in equation (E.7) factorizes as

$$A_t \;=\; \big(1-\mathcal{I}^{-1}_{t-1}\big)\mathbb{E}_{\mathrm{P}_t}\big[s(Y_t, \vartheta_{t|t-1})\big]^2,$$

This term is negative if $1 < \mathcal{I}^{-1}_{t-1}$, equivalently $1 > \mathcal{I}_{t-1}$. That is when the fisher information scalar is small enough the first order term is sufficient for ϕ_1 to be P_t–contractively dominate ϕ_2. Hence, when the Fisher information is sufficiently small, the first-order contribution favors the inverse Fisher scaled update ϕ_1 over the identity scaled update ϕ_2.

This admits a natural interpretation. The Fisher information measures the local variability, or sensitivity, of the score. When $\mathcal{I}_{t-1}$ is small, the score is relatively stable and contains less local curvature information. In this regime the inverse Fisher scaling satisfies

$$\mathcal{I}^{-1}_{t-1} > 1,$$

so the update is amplified relative to the identity scaling. The procedure therefore takes a larger step in the average score direction, allowing the parameter to move more aggressively toward the true value.

Conversely, when $\mathcal{I}_{t-1} > 1$, the score exhibits larger local variability. The inverse Fisher scaling then shrinks the update, damping the step size to avoid overly aggressive movements in a region where the score is highly sensitive to perturbations in the parameter.

We can use the second order term to refine this interpretation.

Second-order term. By Assumption 1(iii), the Hessian

$$H_{\mathrm{LogS}}(Y_t, \vartheta) \;\coloneqq\; \nabla^2_{\vartheta} \mathrm{LogS}(\mathrm{F}_{\vartheta}, Y_t) \;\leq\; 0$$

for every $\vartheta \in \Theta$ and P_t-almost every Y_t.

On the event $\{\mathcal{I}_{t-1} \geq 1\}$ we have

$$0 < \mathcal{I}_{t-1}^{-2} \leq 1.$$

Furthermore,

$$\vartheta^{(1)}_{t|t} = \vartheta_{t|t-1} + \alpha \mathcal{I}_{t-1}^{-1} s\big(\check{Y}_t, \vartheta_{t|t-1}\big), \qquad \vartheta^{(2)}_{t|t} = \vartheta_{t|t-1} + \alpha s\big(\check{Y}_t, \vartheta_{t|t-1}\big),$$

so the inverse Fisher scaling shrinks the update relative to the identity scaling.

Using the definition of the integrated Hessian,

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_t}\Big[\mathcal{I}_{t-1}^{-2}\, s\big(\check{Y}_t, \vartheta_{t|t-1}\big)^{\,2}\, \widetilde{H}_{\mathrm{LogS}}(Y_t, \vartheta^{(1)}_{t|t})\Big] \\
&= \mathbb{E}_{\mathrm{P}_t}\left[\mathcal{I}_{t-1}^{-2}\, s\big(\check{Y}_t, \vartheta_{t|t-1}\big)^{\,2} \int_0^1 (1-u)\, H_{\mathrm{LogS}}\big(Y_t, \vartheta_{t|t-1} + u\alpha \mathcal{I}_{t-1}^{-1} s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\big)\, \mathrm{d}u\right].
\end{aligned}$$

Applying the change of variables

$$z = u\mathcal{I}_{t-1}^{-1}, \qquad du = \mathcal{I}_{t-1}\, dz,$$

gives

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_t}\Big[\mathcal{I}_{t-1}^{-2}\, s\big(\check{Y}_t, \vartheta_{t|t-1}\big)^{\,2}\, \widetilde{H}_{\mathrm{LogS}}(Y_t, \vartheta_{t|t}^{(1)})\Big] \\
&= \mathbb{E}_{\mathrm{P}_t}\Bigg[s\big(\check{Y}_t, \vartheta_{t|t-1}\big)^{\,2} \int_0^{\mathcal{I}_{t-1}^{-1}} (\mathcal{I}_{t-1}^{-1} - z)\, H_{\mathrm{LogS}}\big(Y_t, \vartheta_{t|t-1} + \alpha z s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\big)\, \mathrm{d}z\Bigg].
\end{aligned}$$

Since $\mathcal{I}_{t-1}^{-1} \leq 1$ adapted to $\mathcal{F}_{t-1}^{Y}$, for every $z \in [0, \mathcal{I}_{t-1}^{-1}]$,

$$0 \leq \mathcal{I}_{t-1}^{-1} - z \leq 1 - z.$$

Moreover, by Assumption 1(iii),

$$E_{\mathrm{P}_t}\Big[H_S(Y_t^k, \vartheta)\Big] \leq 0 \quad \text{for all } \vartheta \in \Theta,$$

so multiplying by the larger weight $(1 - z)$ makes the integral more negative. Hence

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_t}\Bigg[s\big(\check{Y}_t, \vartheta_{t|t-1}\big)^{\,2} \int_0^{\mathcal{I}_{t-1}^{-1}} (\mathcal{I}_{t-1}^{-1} - z)\, H_{\mathrm{LogS}}\big(Y_t, \vartheta_{t|t-1} + \alpha z s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\big)\, \mathrm{d}z\Bigg] \\
&\geq \mathbb{E}_{\mathrm{P}_t}\Bigg[s\big(\check{Y}_t, \vartheta_{t|t-1}\big)^{\,2} \int_0^{\mathcal{I}_{t-1}^{-1}} (1 - z)\, H_{\mathrm{LogS}}\big(Y_t, \vartheta_{t|t-1} + \alpha z s\big(\check{Y}_t, \vartheta_{t|t-1}\big)\big)\, \mathrm{d}z\Bigg].
\end{aligned}$$

Since the integrand is nonpositive, extending the integration domain from $[0, \mathcal{I}_{t-1}^{-1}]$ to

$[0,1]$ can only decrease the value of the integral. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_t}\Bigg[s\big(\check{Y}_t,\vartheta_{t|t-1}\big)^{\,2}\int_0^{\mathcal{I}_{t-1}^{-1}}(1-z)\,H_{\mathrm{LogS}}\big(Y_t,\vartheta_{t|t-1}+\alpha z s\big(\check{Y}_t,\vartheta_{t|t-1}\big)\big)\,\mathrm{d}z\Bigg]\\
&\geq \mathbb{E}_{\mathrm{P}_t}\Bigg[s\big(\check{Y}_t,\vartheta_{t|t-1}\big)^{\,2}\int_0^{1}(1-z)\,H_{\mathrm{LogS}}\big(Y_t,\vartheta_{t|t-1}+\alpha z s\big(\check{Y}_t,\vartheta_{t|t-1}\big)\big)\,\mathrm{d}z\Bigg]\\
&= \mathbb{E}_{\mathrm{P}_t}\Big[s\big(\check{Y}_t,\vartheta_{t|t-1}\big)^{\,2}\,\widetilde{H}_{\mathrm{LogS}}(Y_t,\vartheta_{t|t}^{(2)})\Big].
\end{aligned}$$

Combining the inequalities yields

$$\mathbb{E}_{\mathrm{P}_t}\Big[\mathcal{I}_{t-1}^{-2}\,s\big(\check{Y}_t,\vartheta_{t|t-1}\big)^{\,2}\,\widetilde{H}_{\mathrm{LogS}}(Y_t,\vartheta_{t|t}^{(1)})\Big] \geq \mathbb{E}_{\mathrm{P}_t}\Big[s\big(\check{Y}_t,\vartheta_{t|t-1}\big)^{\,2}\,\widetilde{H}_{\mathrm{LogS}}(Y_t,\vartheta_{t|t}^{(2)})\Big].$$

Taking expectations over $\check{Y}_t$ under P_t gives $Q_t \geq 0$ on $\{\mathcal{I}_{t-1} \geq 1\}$.

The argument on the event $\{\mathcal{I}_{t-1} \leq 1\}$ proceeds symmetrically, yielding $Q_t \leq 0$.

Combined ranking. Combining the first- and second-order terms yields

$$\Delta\Delta\mathbb{D}_{\mathrm{LogS}}^{\mathrm{P}_t}\big(\phi_1 \,\|\, \phi_2\big) = -\alpha A_t + \alpha^2 Q_t.$$

When $\mathcal{I}_{t-1} < 1$, the inverse Fisher scaling satisfies $\mathcal{I}_{t-1}^{-1} > 1$, so the score update is amplified relative to the identity scaling. In this regime the score is locally less variable, meaning that a larger step in the score direction is beneficial. Consequently, $A_t < 0$ and $-\alpha A_t > 0$, so the first-order contribution favors the inverse Fisher scaled update ϕ_1.

Conversely, when $\mathcal{I}_{t-1} > 1$, the inverse Fisher scaling shrinks the score update since $\mathcal{I}_{t-1}^{-1} < 1$. Here the score is locally more variable, so aggressive updates become less reliable. The Fisher scaling then acts as a stabilizer by damping the update magnitude. In this case $A_t > 0$ and $-\alpha A_t < 0$ meaning that the first-order term favors the identity scaled update ϕ_2.

The second-order term Q_t captures the curvature effect of the logarithmic scoring rule

along the update path. Since the expected hessian of the logarithmic score is negative assumed to be negative analogue to Assumption 1, larger updates incur a larger curvature penalty. This explains why the sign of Q_t is opposite to that of A_t. Indeed, when $\mathcal{I}_{t-1} < 1$, the inverse Fisher scaling expands the update and therefore traverses a longer path through the negatively curved geometry of the score. This produces a larger negative curvature contribution, yielding $Q_t \leq 0$. Conversely, when $\mathcal{I}_{t-1} > 1$, the inverse Fisher scaling shrinks the update and reduces the curvature penalty relative to the identity scaling, implying $Q_t \geq 0$.

Hence the two terms admit a natural interpretation; the first-order term rewards aggressive movement toward the true parameter, whereas the second-order term penalizes excessively large steps through the local curvature of the scoring rule. The inverse Fisher scaling balances these effects automatically through the local magnitude of the Fisher information.

Example E.1 is the extended version of Example 3 in the main text, providing the full details of the calculation.

Example E.1. *Consider the setting of Example 3. With $\vartheta_{t|t-1} = 1$ and $\mathrm{P}_t = \mathcal{N}(0, 2)$, the contraction dominance criterion reduces to*

$$\Delta\Delta\mathbb{D}^{\mathrm{P}_t}_{\mathrm{LogS}}(\phi_1 \| \phi_2) = \mathbb{E}_{\mathrm{P}_t^{\otimes 2}}\left[\mathrm{LogS}\big(\phi_1(\check{Y}_t, 1), Y_t\big) - \mathrm{LogS}\big(\phi_2(\check{Y}_t, 1), Y_t\big)\right],$$

where $\phi_i(\check{y}, 1) = 1 + \alpha\mathcal{S}^{(i)}_{t-1} s(\check{y}, 1)$ and $s(\check{y}, 1) = \frac{1}{2}(\check{y}^2 - 1)$. Since $Y_t \perp \check{Y}_t$, the inner expectation over Y_t evaluates to $\overline{S}_{\mathrm{LogS}}(\vartheta', \mathrm{P}_t) = -\frac{1}{2}\log(2\pi\vartheta') - 1/\vartheta'$, leaving the single integral

$$\Delta\Delta\mathbb{D}^{\mathrm{P}_t}_{\mathrm{LogS}}(\phi_1 \| \phi_2) = \int_{-\infty}^{\infty} \left[-\frac{1}{2}\log\frac{1 + \alpha(\check{y}^2 - 1)}{1 + \frac{\alpha}{2}(\check{y}^2 - 1)} - \frac{1}{1 + \alpha(\check{y}^2 - 1)} + \frac{1}{1 + \frac{\alpha}{2}(\check{y}^2 - 1)}\right] \frac{e^{-\check{y}^2/4}}{\sqrt{4\pi}}\, d\check{y}.$$

Setting this equal to zero yields $\alpha^* = 0.109$. *The individual divergence reductions* $\Delta\mathbb{D}^{\mathrm{P}_t}_{\mathrm{LogS}}(\phi_i)$, *obtained by replacing* $\mathrm{LogS}(\phi_j(\check{Y}_t, 1), Y_t)$ *with* $\mathrm{LogS}(\vartheta_{t|t-1}, Y_t)$ *in the second term, are evaluated analogously, giving PRADA bounds* $\bar{\alpha}_1 = 0.195$ *and* $\bar{\alpha}_2 = 0.390$ *for* ϕ_1 *and* ϕ_2 *respectively. The binding bound* $\bar{\alpha} = 0.195$ *confirms that the dominance interval* $(0, 0.109)$ *is non-trivially contained within the PRADA interval* $(0, 0.195)$.

For interpretation, the expansion $-\alpha A_t + \alpha^2 Q_t(\alpha)$ *from Theorem 1 identifies* $A_t = -\frac{1}{4} < 0$ *as the source of the first-order gain: small Fisher information allows the step to be amplified beyond identity scaling. The threshold* $\alpha = 0.109$ *marks the point at which the path-integral curvature penalty* $Q_t(\alpha)$ *exactly offsets this gain.*