ONLINE INFERENCE WITH MOMENTUM AND NONSMOOTH SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Momentum changes the transient behavior of stochastic gradient algorithms. We ask when these changes preserve statistical inference after Polyak-Ruppert averaging. We develop a framework for fixed stable score filters that includes SGD, fixed preconditioning, scalar and coordinate momentum, and higher order filters. Under explicit local and trajectory conditions, the filter gain cancels from the first order influence function. Averaged estimators therefore share the usual covariance for the same estimating equation and retained observations. The analysis permits nondifferentiable or discontinuous sample scores by imposing differentiability on their population mean and local quadratic mean continuity. We construct one pass intervals, including an online Jacobian estimator for quantile regression that leaves the nonsmooth update unchanged. An integrated L^2 process limit also justifies path based self-normalization without the additional maximal boundary condition used here to establish uniform path convergence. Simulations show lower error for momentum in one least squares setting with $d = 100$, and online quantile coverage near the nominal level across the reported regular settings.

1 INTRODUCTION

Stochastic gradient descent (SGD) processes observations sequentially with modest computational and memory requirements. Momentum and related score filters can improve its transient behavior, but an optimization improvement does not by itself justify confidence intervals for the estimated population parameter. Polyak-Ruppert averaging connects recursive estimation with statistical inference: under suitable conditions, the averaged iterate has a root-n limit with the usual covariance (Polyak & Juditsky, 1992). We ask whether fixed dynamic filtering preserves this guarantee, including when sample scores are nonsmooth. In smooth strongly convex settings, Tang et al. (2024) establish asymptotic normality and first-order equivalence of averaged scalar SGDM and SGD; we take this equivalence as a starting point and study inference for fixed finite-dimensional score filters when sample scores may be nondifferentiable or discontinuous, formulating the analysis through the population estimating equation and separating local score regularity from the trajectory conditions required for averaging.

Let $\theta_\star$ solve $g(\theta_\star) = 0$, where $g(\theta) = \mathbb{E}\{G(\theta, Z)\}$ has nonsingular Jacobian A at the target and $\mathrm{S} = \mathrm{Var}\{G(\theta_\star, Z)\}$. Classical averaging yields covariance $\Sigma = \mathrm{A}^{-1}\mathrm{S}\mathrm{A}^{-\mathsf{T}}$ under appropriate regularity and stepsizes (Robbins & Monro, 1951; Ruppert, 1988; Kushner & Yang, 1993). Fixed finite-dimensional filters transform both the local drift and the score noise, and their long-run gain need not equal the identity. An exact change of variables shows why a common nonsingular gain cancels from the influence function; the proof then controls the filter state, boundary terms and nonlinear remainder.

Nonsmooth scores create an additional inferential difficulty: differentiating the quantile indicator score does not recover the population Jacobian. We therefore impose differentiability on the population mean and local quadratic-mean continuity on the centered score, without smoothing the update. Path based inference additionally requires joint convergence of the partial sum path and its endpoint, which we establish using an L^2 path limit.

First, we establish a common asymptotic linear representation for Polyak-Ruppert averages of fixed finite-dimensional filtered recursions, permitting discontinuous sample scores through differentiability of the population mean and local quadratic mean score continuity. Under explicit localization, rate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and remainder conditions, the leading influence function does not depend on the filter for a given estimating equation and retained sample; separate results for scalar momentum give sufficient conditions for the required trajectory properties.

Second, we separate covariance consistency from the limit of the averaged estimator. For quantile regression, a residual kernel Jacobian estimator evaluated at predictable iterates is consistent while the original indicator score update is retained; under the stated density envelope and weighted moment conditions, the parameter error need not be negligible relative to the bandwidth.

Third, joint convergence of the partial-sum path in $L^2([0,1])$ and its terminal value is enough for integrated self-normalization, extending path-based inference to fixed filtered recursions without the additional maximal control required for a uniform path limit.

1.1 Related work

Classical stochastic approximation and averaging theory provide the basis for the limit (Polyak & Juditsky, 1992; Kushner & Yin, 2003). Feasible methods include online bootstrap (Fang et al., 2018), recursive asymptotic variance estimation (Godichon-Baggioni, 2019), plug-in and batch means estimators (Chen et al., 2020), solution path random scaling (Lee et al., 2022), and fully online iterate based covariance estimation (Zhu et al., 2023). We ask when such inference remains valid after a fixed dynamic transformation of the stochastic score.

Existing momentum results establish several forms of equivalence with SGD. Yuan et al. (2016) compare mean square behavior in a small constant stepsize regime after rescaling the gain. Heavy-ball convergence and limit laws are studied by Gadat et al. (2018); variable sample size accelerated methods admit central limit theorems under their respective assumptions (Lei & Shanbhag, 2025). For averaged linear stochastic approximation with a constant stepsize, Mou et al. (2020) separate the classical covariance from a stepsize dependent correction and show that momentum leaves the classical term unchanged.

Nonsmooth online inference and self-normalization have established precedents. Online bootstrap covers M-estimation and quantile regression (Fang et al., 2018). Davis et al. (2024) develop asymptotic normality and local optimality through an active manifold framework, while Jiang et al. (2025) analyze online batch means covariance estimation for nonsmooth stochastic approximation. Lee et al. (2025) combine quantile subgradient descent, averaging and random scaling. We extend these inferential conclusions to fixed filtered estimating equations, provide a direct online quantile Jacobian, and use an L^2 path limit tailored to the integrated normalizer.

Sections 2 and 3 present the method and inferential theory; numerical comparisons and concluding remarks follow. The appendices contain trajectory sufficient conditions, model specializations, proofs, simulation details and a power-plant quantile illustration.

2 Methodology

Let $Z_1, Z_2, \ldots$ be independent and identically distributed observations on a measurable space $(\mathcal{Z}, \mathcal{A})$. For a fixed parameter dimension d, let $G : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ be a measurable estimating function with $\mathbb{E}\|G(\theta, Z_1)\| < \infty$ for each $\theta \in \mathbb{R}^d$. Define the population score $g(\theta) = \mathbb{E}\{G(\theta, Z_1)\}$ and the target $\theta_\star \in \mathbb{R}^d$ by $g(\theta_\star) = 0$. We construct an averaged recursive estimator and confidence intervals for prespecified linear contrasts of $\theta_\star$; the construction accommodates smooth, nondifferentiable and discontinuous sample scores. We use the standard O_p and o_p notation; $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution.

Filtered updates and averaging. Let $v_t \in \mathbb{R}^r$ be the filter state at iteration t, with d and r fixed, and let $\mathrm{B} \in \mathbb{R}^{r\times r}$, $\mathrm{C} \in \mathbb{R}^{r\times d}$ and $\mathrm{P} \in \mathbb{R}^{d\times r}$ be fixed matrices. Starting from square-integrable initial values θ_0 and v_0, independent of the observations, we use

$$\begin{aligned} G_t &= G(\theta_{t-1}, Z_t), \qquad v_t = \mathrm{B}v_{t-1} + \mathrm{C}G_t, \\ \theta_t &= \theta_{t-1} - \eta_t \mathrm{P}v_t, \quad \eta_t = \eta_0 t^{-\alpha}, \qquad t \geq 1, \end{aligned} \tag{1}$$

where $\eta_0 > 0$ and $0 < \alpha < 1$. Let $\mathcal{F}_t = \sigma(\theta_0, v_0, Z_1, \ldots, Z_t)$, so that θ_{t-1} is $\mathcal{F}_{t-1}$-measurable and Z_t is independent of $\mathcal{F}_{t-1}$; I_d is the $d \times d$ identity matrix.

Assumption 2.1. *The spectral radius of* B *satisfies* $\rho(\mathrm{B}) < 1$*, and the long-run gain* $\mathrm{K} = \mathrm{P}(\mathrm{I}_r - \mathrm{B})^{-1}\mathrm{C} \in \mathbb{R}^{d\times d}$ *is nonsingular.*

Here $\mathrm{K} = \sum_{j\geq 0} \mathrm{P}\mathrm{B}^j\mathrm{C}$ is the cumulative score response. With $r = d$, SGD takes $\mathrm{B} = 0$, $\mathrm{C} = \mathrm{P} = \mathrm{I}_d$; nonsingular P gives fixed preconditioning. Scalar momentum uses $\mathrm{B} = \beta\mathrm{I}_d$, $\mathrm{C} = c\mathrm{I}_d$, with $0 \leq \beta < 1$, $c > 0$ and gain $\mathrm{K} = c\mathrm{P}/(1-\beta)$. Normalized and unnormalized momentum correspond to $c = 1-\beta$ and $c = 1$. Diagonal B with $\mathrm{C} = \mathrm{I}_d$ gives coordinate momentum, and enlarged states represent finite-order filters. Parameter convergence additionally requires the trajectory conditions in Section 3.

For a sample of size n, let b_n be a deterministic integer burn-in. Define the retained sample size $N_n = n - b_n$ and the Polyak-Ruppert average $\bar{\theta}_{n,b} = N_n^{-1}\sum_{t=b_n+1}^{n}\theta_t$. We impose

$$0 \leq b_n < n, \qquad N_n \longrightarrow \infty, \qquad \limsup_{n\to\infty} \frac{b_n}{n} < 1. \tag{2}$$

Condition (2) retains a positive fraction of the iterates; the full average is $\bar{\theta}_n = n^{-1}\sum_{t=1}^{n}\theta_t$. Under the assumptions of Theorem 3.4, a fixed burn-in or $b_n = o(n)$ gives the same first-order expansion as the full average, and for a positive-fraction burn-in the inferential scale is $\sqrt{N_n}$. Section 3 gives the conditions for the limits below.

From filtered updates to an influence function. Let $\mathrm{A} = \partial g(\theta_\star)/\partial\theta^{\mathsf{T}} \in \mathbb{R}^{d\times d}$ denote the nonsingular Jacobian of the population score. Define its mean remainder by

$$g(\theta_\star + e) = \mathrm{A}e + \mathfrak{r}(e),$$

for every $e \in \mathbb{R}^d$. Local differentiability gives $\|\mathfrak{r}(e)\| = o(\|e\|)$ as $e \to 0$. This expansion concerns the population score and does not require differentiability of individual sample scores. Define the centered score $\xi(\theta, z) = G(\theta, z) - g(\theta)$, its realization $\xi_t(\theta) = \xi(\theta, Z_t)$, and the target score $\xi_t^\star = \xi_t(\theta_\star) = G(\theta_\star, Z_t)$. Assume $\mathbb{E}\|\xi_1^\star\|^2 < \infty$, and write $\mathrm{S} = \mathbb{E}(\xi_1^\star\xi_1^{\star\mathsf{T}})$.

Set $e_t = \theta_t - \theta_\star$, $\mathrm{D} = \mathrm{P}\mathrm{B}(\mathrm{I}_r - \mathrm{B})^{-1} \in \mathbb{R}^{d\times r}$, and $u_t = e_t - \eta_t\mathrm{D}v_t$. At $t = 0$, use $\eta_0 = \eta_1$ in this definition. Since $\mathrm{P} + \mathrm{D} = \mathrm{P}(\mathrm{I}_r - \mathrm{B})^{-1}$, substitution into (1) gives the exact identity

$$u_t = u_{t-1} - \eta_t\mathrm{K}G_t + (\eta_{t-1} - \eta_t)\mathrm{D}v_{t-1}. \tag{3}$$

Substituting the exact score decomposition $G_t = \mathrm{A}e_{t-1} + \xi_t^\star + \mathfrak{r}(e_{t-1}) + \{\xi_t(\theta_{t-1}) - \xi_t^\star\}$ into (3) shows that the same gain K acts on both drift and noise, giving the influence term $-(\mathrm{K}\mathrm{A})^{-1}\mathrm{K}\xi_t^\star = -\mathrm{A}^{-1}\xi_t^\star$. Rearranging the recursion expresses $\mathrm{A}e_{t-1}$ as $-\xi_t^\star$ plus weighted increments of u_t and remainder terms; under Assumptions 3.1 to 3.3 and (2), these additional terms sum to $o_p(\sqrt{N_n})$, so that, for the original average $\bar{\theta}_{n,b}$,

$$\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{N_n}}\sum_{t=b_n+1}^{n}\xi_t^\star + o_p(1). \tag{4}$$

Theorem 3.4 establishes (4) and $\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_\star) \xrightarrow{d} \mathcal{N}_d(0, \Sigma)$, where $\Sigma = \mathrm{A}^{-1}\mathrm{S}\mathrm{A}^{-\mathsf{T}}$. Under the stated trajectory conditions, admissible filters therefore share the same first-order covariance for the same score and retained observations.

Online inference. By (4), standard errors need A and S. Estimate S from the unfiltered scores by $\widehat{\mathrm{S}}_{n,b} = N_n^{-1}\sum_{t=b_n+1}^{n} G_tG_t^{\mathsf{T}}$, let $\widehat{\mathrm{A}}_n \in \mathbb{R}^{d\times d}$ be an online estimator of A and, since A need not be symmetric, define

$$\mathrm{R}_n = (\widehat{\mathrm{A}}_n^{\mathsf{T}}\widehat{\mathrm{A}}_n + \lambda_n\mathrm{I}_d)^{-1}\widehat{\mathrm{A}}_n^{\mathsf{T}}, \qquad \widehat{\Sigma}_{n,b} = \mathrm{R}_n\widehat{\mathrm{S}}_{n,b}\mathrm{R}_n^{\mathsf{T}}, \tag{5}$$

where deterministic $\lambda_n > 0$ tends to zero. The ridge keeps R_n well defined at every iteration and leaves the parameter update unchanged. Under Assumption 3.5, $\widehat{\mathrm{A}}_n \xrightarrow{p} \mathrm{A}$ and $\widehat{\mathrm{S}}_{n,b} \xrightarrow{p} \mathrm{S}$, so continuity of (5) gives $\mathrm{R}_n \xrightarrow{p} \mathrm{A}^{-1}$ and $\widehat{\Sigma}_{n,b} \xrightarrow{p} \Sigma$.

The unfiltered scores in $\widehat{\mathrm{S}}_{n,b}$ correspond to the independent target-score innovations in (4). A filter state combines several past scores, so successive states share inputs and their contemporaneous

second moment generally differs from S. The cancellation of K therefore does not justify replacing $G_t G_t^\mathsf{T}$ by $v_t v_t^\mathsf{T}$. Moreover, evaluating scores and Jacobian contributions at θ_{t-1} keeps the evaluation point determined by past observations. This is the predictable structure used when conditioning on $\mathcal{F}_{t-1}$ in the covariance arguments.

Let $a \in \mathbb{R}^d$ be a prespecified contrast with $a^\mathsf{T}\Sigma a > 0$ and $\gamma \in (0,1)$ the significance level. Under the assumptions of Theorems 3.4 and 3.6, Slutsky's theorem gives $\sqrt{N_n}\, a^\mathsf{T}(\bar{\theta}_{n,b} - \theta_\star)/(a^\mathsf{T}\widehat{\Sigma}_{n,b}a)^{1/2} \xrightarrow{d} \mathcal{N}(0,1)$, so the confidence interval

$$a^\mathsf{T}\bar{\theta}_{n,b} \;\pm\; z_{1-\gamma/2}\left(\frac{a^\mathsf{T}\widehat{\Sigma}_{n,b}a}{N_n}\right)^{1/2}, \tag{6}$$

where $z_{1-\gamma/2}$ is the standard normal $(1-\gamma/2)$ quantile, has limiting coverage $1-\gamma$.

For the Jacobian constructions below, let $\mathrm{H}_t \in \mathbb{R}^{d\times d}$ be the contribution from (θ_{t-1}, Z_t) and set $\widehat{\mathrm{A}}_n = n^{-1}\sum_{t=1}^n \mathrm{H}_t$, updated online as $\widehat{\mathrm{A}}_t = \widehat{\mathrm{A}}_{t-1} + t^{-1}(\mathrm{H}_t - \widehat{\mathrm{A}}_{t-1})$ from $\widehat{\mathrm{A}}_0 = 0$ using all observations, including burn-in. For differentiable scores, $\mathrm{H}_t = D_\theta G(\theta_{t-1}, Z_t)$ is the sample-score Jacobian; this requires interchanging differentiation and expectation, together with the moment and continuity conditions that give $\widehat{\mathrm{A}}_n \xrightarrow{p} \mathrm{A}$. Section C.4 verifies concrete constructions for the regression models.

Quantile regression illustrates a Jacobian construction for a discontinuous score. Write $Z = (X, Y)$ with regressor $X \in \mathbb{R}^d$ and response $Y \in \mathbb{R}$, and $Z_t = (X_t, Y_t)$. For a fixed quantile level $\tau \in (0,1)$, use $G(\theta; X, Y) = X\{\mathbf{1}(Y \le X^\mathsf{T}\theta) - \tau\}$, where $\mathbf{1}(\cdot)$ is the indicator. Let $f(y \mid X)$ denote the conditional density of Y given X. Under the quantile regularity conditions in Section C.4, the population Jacobian is $\mathrm{A} = \mathbb{E}\{f(X^\mathsf{T}\theta_\star \mid X)XX^\mathsf{T}\}$. Differentiating the indicator score does not recover this matrix. Instead, let $\mathcal{K} : \mathbb{R} \to [0,\infty)$ be a bounded, compactly supported kernel with $\int \mathcal{K}(u)\,\mathrm{d}u = 1$. For $h_t = h_0 t^{-\kappa}$, with $h_0 > 0$ and $0 < \kappa < 1$, use

$$\mathrm{H}_t = \frac{1}{h_t}\mathcal{K}\left(\frac{Y_t - X_t^\mathsf{T}\theta_{t-1}}{h_t}\right)X_t X_t^\mathsf{T}. \tag{7}$$

The average of (7) is the estimator in (16), consistent under the additional density, weighted-moment and trajectory conditions of Theorem 3.11. Smoothing is confined to Jacobian estimation; the parameter recursion retains the original indicator score. Since the kernel contributions are positive semidefinite and the limiting Jacobian is positive definite, Theorem 3.11 also establishes the direct ridge inverse $(\widehat{\mathrm{A}}_n + \lambda_n \mathrm{I}_d)^{-1}$; Algorithm 1 uses the general inverse in (5).

Self-normalized inference. An alternative uses the trajectory of the full average and avoids estimation of A and S. Following the solution-path construction of Lee et al. (2022), define

$$\widehat{\mathrm{V}}_n = \frac{1}{n^2}\sum_{k=1}^n k^2(\bar{\theta}_k - \bar{\theta}_n)(\bar{\theta}_k - \bar{\theta}_n)^\mathsf{T}. \tag{8}$$

The integrated process limit in Theorem 3.7 justifies inference based on (8) for the filtered recursion: in each scalar contrast, the same asymptotic scale multiplies the endpoint error and its path fluctuations, and their ratio removes it. For each positive integer q, let $W_q(s)$, $0 \le s \le 1$, denote standard q-dimensional Brownian motion, and define $\mathrm{J}_q = \int_0^1 \{W_q(s) - sW_q(1)\}\{W_q(s) - sW_q(1)\}^\mathsf{T}\,\mathrm{d}s$. Under the assumptions of Theorem 3.8, the scalar pivot $\sqrt{n}\, a^\mathsf{T}(\bar{\theta}_n - \theta_\star)/(a^\mathsf{T}\widehat{\mathrm{V}}_n a)^{1/2}$ converges in distribution to $\mathcal{T}_1 = W_1(1)/\sqrt{\mathrm{J}_1}$. Let $c^{\mathrm{SN}}_{1-\gamma}$ be the $(1-\gamma)$ quantile of $|\mathcal{T}_1|$. The corresponding interval is

$$a^\mathsf{T}\bar{\theta}_n \;\pm\; c^{\mathrm{SN}}_{1-\gamma}\left(\frac{a^\mathsf{T}\widehat{\mathrm{V}}_n a}{n}\right)^{1/2}. \tag{9}$$

Interval (9) has limiting coverage $1-\gamma$; its center and normalizer use the full path from iteration 1, and the reference distribution contains no unknown model parameters, so its critical values are computed once. Both interval constructions give asymptotic coverage at prespecified reporting horizons. For joint inference, let $\mathrm{R}_0 \in \mathbb{R}^{q\times d}$ have rank q, $1 \le q \le d$, let $r_0 \in \mathbb{R}^q$ specify the null hypothesis $\mathrm{R}_0\theta_\star = r_0$, and suppose $\mathrm{R}_0\Sigma\mathrm{R}_0^\mathsf{T}$ is positive definite. The self-normalized test rejects when $n(\mathrm{R}_0\bar{\theta}_n - r_0)^\mathsf{T}(\mathrm{R}_0\widehat{\mathrm{V}}_n\mathrm{R}_0^\mathsf{T})^{-1}(\mathrm{R}_0\bar{\theta}_n - r_0)$ exceeds the $(1-\gamma)$ quantile of its limit $W_q(1)^\mathsf{T}\mathrm{J}_q^{-1}W_q(1)$ (Theorem 3.8); the sample normalizer is nonsingular with probability tending to one.

One-pass implementation. Algorithm 1 in Appendix C.6 gives the parameter, covariance and self-normalizer updates. Scores and Jacobian contributions are evaluated before each parameter update. With a fixed burn-in, running accumulators compute the retained average, the online inference and the full-path self-normalizer without storing observations or iterates. Dense updates require $O(d^2 + r^2)$ memory and $O(d^2 + r^2 + dr)$ arithmetic, apart from score and Jacobian evaluation and reporting factorizations; q self-normalized contrasts add $O(q^2)$ storage. Burn-in affects the retained mean and score second moment, but observations during burn-in still update the parameter, filter state and Jacobian estimate. For a prespecified terminal horizon n, a fractional burn-in can be implemented by fixing $b = \lfloor \rho n \rfloor$ in advance, with $0 \le \rho < 1$. The same retained accumulators then apply throughout that run. Full-path self-normalization requires its own accumulators from the first iteration even when a tail-average Wald interval is reported alongside it. Equation (23) computes this normalizer from running sums, so the two inference branches can share the parameter trajectory while retaining their respective averaging conventions.

3 Theoretical results

3.1 Common influence function and feasible inference

Write $e_t = \theta_t - \theta_\star$. The assumptions separate local properties of the score from those of the computed trajectory.

Assumption 3.1. *For some neighborhood $\mathcal{U}$ of $\theta_\star$, $g(\theta_\star + e) = \mathrm{A}e + \mathfrak{r}(e)$ whenever $\theta_\star + e \in \mathcal{U}$, where A is nonsingular and $\|\mathfrak{r}(e)\| = o(\|e\|)$ as $e \to 0$.*

Assumption 3.2. *The target score satisfies $\mathbb{E}\|\xi_1^\star\|^2 < \infty$. There exists a bounded nondecreasing function $\omega_\xi : [0,\infty) \to [0,\infty)$ with $\omega_\xi(0+) = 0$ such that $\mathbb{E}\|\xi(\theta, Z_1) - \xi(\theta_\star, Z_1)\|^2 \le \omega_\xi(\|\theta - \theta_\star\|)$ for all $\theta \in \mathcal{U}$.*

Assumption 3.3. *Every computed score G_t is finite almost surely. With $\eta_t = \eta_0 t^{-\alpha}$ and $0 < \alpha < 1$, the trajectory is eventually confined to $\mathcal{U}$ in the sense that $\mathbb{P}(\theta_t \in \mathcal{U} \text{ for every } t \ge m) \to 1$ as $m \to \infty$. There are constants $L_e < \infty$ and $\varrho > \max\{0, 2\alpha - 1\}$ such that, for all sufficiently large t,*

$$\mathbb{E}\|e_t\|^2 \le L_e t^{-\varrho}, \qquad \sup_{t \ge 0} \mathbb{E}\|v_t\|^2 < \infty. \tag{10}$$

The accumulated nonlinear remainder satisfies

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \|\mathfrak{r}(e_{t-1})\| \, \mathbf{1}\{\theta_{t-1} \in \mathcal{U}\} \xrightarrow{p} 0. \tag{11}$$

The indicator in (11) restricts the remainder to $\mathcal{U}$. Population differentiability supplies local identification (van der Vaart, 1998, Chapter 5); quadratic-mean continuity permits discontinuous sample scores. The trajectory rate controls boundary terms and the accumulated remainder controls nonlinear bias. A local bound $\|\mathfrak{r}(e)\| \le L\|e\|^{1+\nu}$ and $\varrho(1+\nu) > 1$ suffice for the latter (proposition C.2). The scalar momentum results in Appendix C.2.1 supply trajectory conditions; the theoretical appendix verifies a projected nonsmooth example.

Theorem 3.4. *Suppose Assumptions 2.1 to 3.3 and* (2) *hold. Then*

$$\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_\star) = -\mathrm{A}^{-1} \frac{1}{\sqrt{N_n}} \sum_{t=b_n+1}^{n} \xi_t^\star + o_p(1). \tag{12}$$

Consequently, $\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_\star) \xrightarrow{d} \mathcal{N}_d(0, \Sigma)$, where $\mathcal{N}_d$ denotes the d-dimensional normal distribution and $\Sigma = \mathrm{A}^{-1}\mathrm{S}\mathrm{A}^{-\mathsf{T}}$.

For the same score and averaging window, the influence function is independent of the fixed filter. Filters and batch estimators with the same expansion are first order equivalent (corollary C.3); a full-sample batch comparison additionally requires $b_n = o(n)$. This pointwise result allows substantial error at finite horizons near instability or singular gain. For example, if $b_n/n \to \rho \in (0,1)$, expressing the limit in Theorem 3.4 on the $\sqrt{n}$ scale gives covariance $\Sigma/(1-\rho)$. A fixed positive burn-in fraction thus reduces the leading information retained by the average, even though earlier observations still

influence the recursion. Burn-in may attenuate initial transients at finite horizons, but comparisons with a full-sample batch estimator must account for this change of scale as well as for the choice of filter.

Assumption 3.5. *The online scores satisfy*

$$\mathbb{E}\|G_t\|^2 < \infty \quad \textit{for all } t, \qquad \mathbb{E}\|G(\theta_{t-1}, Z_t) - G(\theta_\star, Z_t)\|^2 = o(1).$$

Moreover, $\widehat{\mathrm{A}}_n \xrightarrow{p} \mathrm{A}$*, and* λ_n *is a positive deterministic sequence converging to zero.*

Theorem 3.6. *Suppose* $g(\theta_\star) = 0$*,* A *is nonsingular,* $\mathbb{E}\|G(\theta_\star, Z_1)\|^2 < \infty$*, and Assumption 3.5 holds. For deterministic* $0 \le b_n < n$ *with* $N_n \to \infty$*,* $\widehat{\mathrm{S}}_{n,b} \xrightarrow{p} \mathrm{S}$*,* $\mathrm{R}_n \xrightarrow{p} \mathrm{A}^{-1}$*, and* $\widehat{\Sigma}_{n,b} \xrightarrow{p} \Sigma$*. If the assumptions of Theorem 3.4 also hold, then, for every fixed* $a \in \mathbb{R}^d$ *with* $a^\mathsf{T}\Sigma a > 0$*,*

$$\frac{\sqrt{N_n}\, a^\mathsf{T}(\bar{\theta}_{n,b} - \theta_\star)}{(a^\mathsf{T}\widehat{\Sigma}_{n,b} a)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

Accordingly, (6) *has asymptotic coverage* $1-\gamma$*. If* Σ *is positive definite, the Wald ellipsoid using* $\widehat{\Sigma}_{n,b}^{-1}$ *and the* $(1-\gamma)$ *quantile of the chi-squared distribution with* d *degrees of freedom also has this coverage. The scalar pivot is evaluated when its estimated variance is positive, and the ellipsoid when* $\widehat{\Sigma}_{n,b}$ *is positive definite. These events have probability tending to one. On their complements, the pivots may be extended arbitrarily for the convergence statements; an implementation records the corresponding inference as unavailable.*

Covariance consistency uses score L^2 convergence and the target second moment; Wald inference also needs the averaged estimator limit. Proofs of the main results are collected in the theoretical appendix.

The global drift and noise bounds in Assumption C.4 give the mean-square rate and almost-sure localization for scalar momentum (Theorem C.5). A local alternative projects onto a compact convex neighborhood of the target; under the stated local drift and remainder bounds, projection changes the update only finitely often and the same inferential limit follows (Theorem C.6). Appendix C.2.1 gives both statements and their sufficient conditions.

3.2 Self-normalization

For the full average, let $\mathbb{U}_n(s) = n^{-1/2}\sum_{t \le \lfloor ns \rfloor} e_t$, $\mathbb{H} = L^2([0,1];\mathbb{R}^d)$ and $\Lambda = -\mathrm{A}^{-1}\mathrm{S}^{1/2}$. Here $\mathrm{S}^{1/2}$ is the symmetric positive-semidefinite square root and W_d is standard Brownian motion.

Theorem 3.7. *Under Assumptions 2.1 to 3.3,* $(\mathbb{U}_n, \mathbb{U}_n(1)) \xrightarrow{d} (\Lambda W_d, \Lambda W_d(1))$ *in* $\mathbb{H} \times \mathbb{R}^d$*.*

Since evaluation at $s = 1$ is not continuous on L^2, the endpoint is a separate coordinate; the joint limit justifies the normalizer $\widehat{\mathrm{V}}_n$ in (8).

Theorem 3.8. *Under the assumptions of Theorem 3.7,*

$$\{\sqrt{n}(\bar{\theta}_n - \theta_\star), \widehat{\mathrm{V}}_n\} \xrightarrow{d} \{\Lambda W_d(1), \Lambda \mathrm{J}_d \Lambda^\mathsf{T}\}.$$

For every fixed $a \in \mathbb{R}^d$ *with* $a^\mathsf{T}\Sigma a > 0$*,*

$$\frac{\sqrt{n}\, a^\mathsf{T}(\bar{\theta}_n - \theta_\star)}{(a^\mathsf{T}\widehat{\mathrm{V}}_n a)^{1/2}} \xrightarrow{d} \mathcal{T}_1, \qquad \mathcal{T}_1 = \frac{W_1(1)}{\left[\int_0^1 \{W_1(s) - sW_1(1)\}^2\, ds\right]^{1/2}}. \tag{13}$$

Consequently, the confidence interval in (9) *has asymptotic coverage* $1-\gamma$ *for* $a^\mathsf{T}\theta_\star$*.*

More generally, let $\mathrm{R}_0 \in \mathbb{R}^{q \times d}$ *have rank* q*, let* $r_0 \in \mathbb{R}^q$ *be specified, and suppose* $\Omega = \mathrm{R}_0 \Sigma \mathrm{R}_0^\mathsf{T}$ *is positive definite. Under the null hypothesis* $\mathrm{R}_0\theta_\star = r_0$*,*

$$n(\mathrm{R}_0\bar{\theta}_n - r_0)^\mathsf{T}(\mathrm{R}_0\widehat{\mathrm{V}}_n\mathrm{R}_0^\mathsf{T})^{-1}(\mathrm{R}_0\bar{\theta}_n - r_0) \xrightarrow{d} W_q(1)^\mathsf{T}\mathrm{J}_q^{-1}W_q(1). \tag{14}$$

The sample matrix $\mathrm{R}_0\widehat{\mathrm{V}}_n\mathrm{R}_0^\mathsf{T}$ *is positive definite with probability tending to one. The scalar ratio is evaluated on* $\{a^\mathsf{T}\widehat{\mathrm{V}}_n a > 0\}$ *and the vector statistic on the event that* $\mathrm{R}_0\widehat{\mathrm{V}}_n\mathrm{R}_0^\mathsf{T}$ *is positive definite. On the complementary events, arbitrary extensions leave the weak limits unchanged; an implementation records the corresponding inference as unavailable.*

These reference laws extend established unfiltered random-scaling constructions (Lee et al., 2022; 2025) and remove Jacobian and score-covariance estimation without the additional maximal condition needed for uniform path convergence (proposition C.8); Algorithm 1 gives the updates.

3.3 A FEASIBLE QUANTILE JACOBIAN

We now establish consistency of the online quantile Jacobian constructed in (7). The additional density and weighted moment conditions below are used for covariance estimation. Wald inference also requires the quantile limit in theorem C.17.

Fix $\tau \in (0,1)$, $G_\tau(\theta; X, Y) = X\{\mathbf{1}(Y \le X^\mathsf{T}\theta) - \tau\}$ and $g_\tau = \mathbb{E}G_\tau$. Let $g_\tau(\theta_\tau) = 0$, $q_\tau(X) = X^\mathsf{T}\theta_\tau$ and $F(y \mid X) = \mathbb{P}(Y \le y \mid X)$; the target permits misspecification.

Assumption 3.9. *The target θ_τ satisfies $g_\tau(\theta_\tau) = 0$, $\mathbb{E}\|X\|^2 < \infty$, and $\mathbb{P}\{Y = q_\tau(X) \mid X\} = 0$ almost surely. There is a nonnegative measurable function $f_\tau(X)$ such that*

$$\left\|\mathbb{E}\left[X\left\{F(q_\tau(X) + X^\mathsf{T}e \mid X) - F(q_\tau(X) \mid X) - f_\tau(X)X^\mathsf{T}e\right\}\right]\right\| = o(\|e\|) \quad (e \to 0). \tag{15}$$

The matrix $\mathrm{A}_\tau = \mathbb{E}\{f_\tau(X)XX^\mathsf{T}\}$ is finite and positive definite.

With the trajectory assumptions, Theorem 3.4 gives the quantile CLT (Theorem C.17). For feasible covariance estimation, choose a bounded nonnegative compactly supported kernel $\mathcal{K}$ integrating to one, with $\int_{\mathbb{R}} \mathcal{K}(u)^p \,\mathrm{d}u < \infty$, $1 \le p \le 2$, and $h_t = h_0 t^{-\kappa}$, $h_0 > 0$, $0 < \kappa < 1$. Averaging the contributions in (7) gives

$$\widehat{\mathrm{A}}_{\tau,n} = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{h_t}\mathcal{K}\left(\frac{Y_t - X_t^\mathsf{T}\theta_{t-1}}{h_t}\right) X_t X_t^\mathsf{T}. \tag{16}$$

Assumption 3.10. *The conditional distribution of Y given X has a density $f(\cdot \mid X)$, with $f_\tau(X) = f(q_\tau(X) \mid X)$. For some $p \in (1,2]$ and $\nu_A \in (0,1]$, there are nonnegative measurable envelopes $\bar{f}(X)$ and $L_A(X)$ such that, almost surely,*

$$\sup_y f(y \mid X) \le \bar{f}(X), \qquad |f(q_\tau(X) + s \mid X) - f(q_\tau(X) \mid X)| \le L_A(X)|s|^{\nu_A} \quad \textit{for all } s \in \mathbb{R}.$$

The weighted moments satisfy $\mathbb{E}\{\bar{f}(X)\|X\|^{2p}\} < \infty$ and $\mathbb{E}[L_A(X)\{\|X\|^2 + \|X\|^{2+\nu_A}\}] < \infty$.

Set $\widetilde{\mathrm{A}}_{\tau,n} = \widehat{\mathrm{A}}_{\tau,n} + \lambda_n \mathrm{I}_d$ and $\widehat{\mathrm{S}}_{\tau,n,b} = N_n^{-1}\sum_{t>b_n}^{n} G_\tau(\theta_{t-1}; X_t, Y_t)G_\tau(\theta_{t-1}; X_t, Y_t)^\mathsf{T}$, with $\Sigma_\tau = \mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}$ and $\mathrm{S}_\tau = \mathbb{E}[G_\tau(\theta_\tau; X, Y)G_\tau(\theta_\tau; X, Y)^\mathsf{T}]$.

Theorem 3.11. *Suppose Assumptions 3.9 and 3.10 hold and the iterates are predictable relative to the fresh observations, with $\theta_t \xrightarrow{p} \theta_\tau$ and $n^{-1}\sum_{t=m+1}^{n}\mathbb{E}\|\theta_{t-1} - \theta_\tau\|^{\nu_A} \to 0$ for some fixed m, with finite almost-sure iterates before m. For deterministic retained windows with $N_n \to \infty$, $\widehat{\mathrm{A}}_{\tau,n} \xrightarrow{p} \mathrm{A}_\tau$ and $\widetilde{\mathrm{A}}_{\tau,n}^{-1} \xrightarrow{p} \mathrm{A}_\tau^{-1}$. The corresponding estimator satisfies $\widehat{\Sigma}_{\tau,n,b} = \widetilde{\mathrm{A}}_{\tau,n}^{-1}\widehat{\mathrm{S}}_{\tau,n,b}\widetilde{\mathrm{A}}_{\tau,n}^{-1} \xrightarrow{p} \Sigma_\tau$.*

The trajectory rate suffices for the displacement-moment condition. Integrating over the fresh response bounds the kernel bias by $C\|\theta_{t-1} - \theta_\tau\|^{\nu_A} + Ch_t^{\nu_A}$, so no restriction comparing the parameter error with the bandwidth is needed under the density envelope. Wald inference also needs the quantile CLT.

Appendix C.4 specializes the limit to smooth M-estimation, least squares, logistic and Huber regression (corollaries C.9 to C.11 and C.14). It also gives online Jacobians (proposition C.15) and projected logistic and quantile cases (corollary C.12 and example C.18). Compared with smooth scalar momentum (Tang et al., 2024), the fixed-filter results cover nonsmooth scores under the stated trajectory conditions without changing the first-order covariance.

The power-plant example (section B and table 6) illustrates online quantile inference and order sensitivity across five supplied observation orders.

4 NUMERICAL EXPERIMENTS

We study two questions: how stable filters affect finite-sample estimation under matched gains, and how online covariance estimation and alternative interval constructions affect coverage. The

experiments comprise eight least-squares or logistic settings and seven quantile settings, with $M = 500$ replications each. Tables report first-coordinate 95% coverage and per-coordinate RMSE, $\{(Md)^{-1}\sum_{m=1}^{M}\|\widehat{\theta}_m - \theta_\star\|^2\}^{1/2}$. The first coordinate is a slope in the smooth models and an intercept in the quantile models.

Shared implementation. In filter comparisons, Proposed denotes our online inference for gain-matched filtered PR averages (Algorithm 1); parentheses identify the filter: normalized momentum, coordinate momentum, or two-pole. SGD is the unfiltered reference. Recursions start at zero and use $\eta_t = \eta_0 t^{-0.600}$. Independent pilot simulations selected the decay exponent, stepsize multiplier, burn-in fraction and horizon multiplier; β and the filter shapes were prespecified. Appendix A.1 gives the selection criterion and Table 3 lists the frozen configurations. The scale η_0 is distinct from the filter's long-run gain K in Assumption 2.1. Normalized momentum and the two-pole cascade have identity long-run gain; the latter uses poles 0.5 and β. Methods within a replication receive the same data. Batch fits use all observations; recursive Wald intervals use averages of the last 70%, while self-normalized intervals use full averages. Thus comparisons with batch fits or self-normalization change the estimator as well as the covariance construction. Additional filter comparisons and implementation details are given in Appendix A.

Recursive Wald standard errors use the retained size $N_n = n - b_n$; batch and self-normalized intervals use n. The simulations use a fixed ridge of 10^{-6}, distinct from the vanishing ridge required by the consistency theorems.

Smooth scores. For least squares, $X \sim N(0, \Gamma)$ with $\Gamma_{jk} = 0.5^{|j-k|}$ and $Y = X^\mathsf{T}\theta_\star + \varepsilon$. The nonzero slope coefficients are $(0.550, -0.450, 0.300)$. Errors are $N(0,1)$, $t_3/\sqrt{3}$, or $\mathrm{Exp}(1) - \log 2$; mean-zero regressors preserve the no-intercept slope target in the last case. Correctly specified logistic responses have probability $\pi(X^\mathsf{T}\theta_\star)$, with π defined in Appendix C.4. For misspecified logistic responses, the generating log odds use the same slope coefficients and add $0.600(X_1^2 - 1)$. The theoretical pseudo-true slope solves $\mathbb{E}[X\{\pi(X^\mathsf{T}\theta_\star) - p_{\mathrm{true}}(X)\}] = 0$, where $p_{\mathrm{true}}(X)$ is the generating success probability. An approximation from 150000 independent reference covariates is reused in all 500 replications; Appendix A.2 gives its construction. Dimensions are 5, 20 or 100, horizons range from 8000 to 200000, and $\beta \in \{0.5, 0.9, 0.99\}$.

Gain-matched recursions have similar RMSE and coverage in most configurations in Table 1, consistent with their common first-order influence function. In asymmetric least squares with $d = 100$, $n = 8000$ and $\beta = 0.990$, Proposed (momentum) lowers RMSE from 3.700×10^{-2} to 1.700×10^{-2} and raises coverage from 0.910 to 0.964. Proposed (two-pole) performs similarly, while the full-sample batch least-squares estimator has lower error. Agreement at the reported precision does not imply identical paths.

Table 1: Smooth models: 95% coverage of the first slope/per-coordinate RMSE. Recursive intervals use online estimates and the last 70% of observations; batch fits use the full sample.

Configuration	SGD	Proposed (momentum)	Proposed (two-pole)	Batch
LS, Gaussian, small	0.960/0.016	0.960/0.016	0.960/0.016	0.960/0.014
LS, t_3, medium	0.950/0.008	0.950/0.008	0.950/0.008	0.954/0.006
LS, asymmetric, high d	0.910/0.037	0.964/0.017	0.964/0.017	0.952/0.015
LS, Gaussian, large n	0.928/0.003	0.928/0.003	0.928/0.003	0.936/0.003
Logit, correct, small	0.948/0.034	0.948/0.034	0.948/0.034	0.954/0.028
Logit, correct, medium	0.954/0.016	0.954/0.016	0.952/0.016	0.948/0.013
Logit, misspecified	0.960/0.015	0.958/0.015	0.958/0.015	0.966/0.013
Logit, misspecified, large n	0.954/0.007	0.954/0.007	0.954/0.007	0.958/0.006

Additional Proposed (coordinate) and raw-gain comparisons appear in Table 4. Using the SGD stepsize for unnormalized momentum changes the effective gain and produces failures or undercoverage in the reported settings. Rescaling by $1 - \beta$ recovers the normalized path exactly (Proposition C.7).

Quantile estimation. We use $Y = X^\top\theta_\tau + \sigma(X)\varepsilon_\tau$, with $X = (1, W^\top)^\top$, $\tau \in \{0.1, 0.5, 0.9\}$, and $\sigma(X) = 1$ or $0.5 + 0.5|X_2|$. For every error distribution, $\varepsilon_\tau = U - F_U^{-1}(\tau)$, where F_U is the distribution function of U and U follows $N(0,1)$, $t_3/\sqrt{3}$, or $\mathrm{Exp}(1)$. Thus $Q_\tau(\varepsilon_\tau) = 0$. The labels Gaussian, t_3, and asymmetric refer to these error distributions. The first four coefficients are $(0.250, 0.550, -0.450, 0.300)$ and the remainder are zero. The median-t_3 heteroscedastic, median-t_3 large-n, and lower-tail asymmetric settings use t_5 covariates; all other quantile settings use Gaussian covariates. Appendix A.3 specifies both constructions. The online Jacobian uses an Epanechnikov kernel and $h_t = t^{-0.300}$.

Across Table 5 in Appendix A.3, recursive coverage ranges from 0.940 to 0.972. Filtering preserves comparable error and calibration across these nonsmooth updates; it does not produce the large error reduction seen in the high-dimensional least-squares example. Batch quantile regression often has lower RMSE, particularly in the upper-tail heteroscedastic setting.

Quantile inference. Table 2 separates covariance estimation from filtering by holding normalized momentum fixed. Proposed (kernel) uses the online Jacobian in (16). Its coverage and the oracle coverage differ by at most 1.200×10^{-2} for the same point estimate. The oracle uses population covariance formulas with independent reference integration where needed, while the feasible estimator evaluates its kernel at recursive residuals. This comparison isolates the practical cost of estimating the Jacobian.

The equal-block diagnostic uses complete blocks of length $\lfloor N_n^{0.600} \rfloor$ and discards the final incomplete block. Its estimate multiplies the sample covariance of block averages by the block length. This growth need not dominate the trajectory's correlation length; it is not the increasing-block procedure of Zhu et al. (2023); Jiang et al. (2025). The absolute Brownian pivot for self-normalization is calibrated independently with 50000 paths on a 600-point grid. Its coverage ranges from 0.778 to 0.998, so the asymptotic result alone does not ensure accurate calibration at these horizons.

Table 2: Interval comparison for normalized momentum: intercept 95% coverage/per-coordinate RMSE. Proposed (kernel), oracle and equal-block intervals use the tail average; self-normalization uses the full average.

Configuration	Proposed (kernel)	Oracle	Equal blocks	Self-normalized
$\tau = 0.1$, Gaussian	0.960/0.018	0.958/0.018	0.190/0.018	0.958/0.035
$\tau = 0.5$, t_3, heteroscedastic	0.970/0.005	0.964/0.005	0.466/0.005	0.926/0.005
$\tau = 0.9$, asymmetric, heteroscedastic	0.948/0.045	0.936/0.045	0.140/0.045	0.778/0.066
$\tau = 0.5$, Gaussian, small	0.968/0.020	0.964/0.020	0.370/0.020	0.894/0.030
$\tau = 0.5$, t_3, large n	0.956/0.002	0.956/0.002	0.432/0.002	0.938/0.002
$\tau = 0.1$, asymmetric	0.946/0.002	0.938/0.002	0.490/0.002	0.998/0.007
$\tau = 0.9$, t_3	0.948/0.024	0.936/0.024	0.164/0.024	0.834/0.040

5 Concluding remarks

Fixed stable score filters preserve the first-order influence function of Polyak-Ruppert averages under the stated trajectory conditions, providing a unified basis for online inference with momentum and other finite-dimensional filters. The framework accommodates smooth and discontinuous scores, separates trajectory stability from inferential validity, and supports both and self-normalized inference. For quantile regression, the proposed online Jacobian estimator enables feasible covariance estimation without smoothing the underlying indicator-score recursion, while the integrated L^2 limit avoids Jacobian and covariance estimation altogether.

At the studied sample sizes, gain-matched filters generally have comparable error and coverage. Momentum has lower error in one pilot-selected least-squares setting with $d = 100$, and online quantile intervals are well calibrated across the reported regular settings. Extending verifiable trajectory conditions beyond scalar momentum, developing finite-sample approximations, and accommodating dependent observations remain for future work.

AI USE STATEMENT

Generative AI tools were used to revise wording and improve the clarity and readability of the manuscript; to identify potentially relevant literature, references, and search directions, and to assist with research execution, including drafting, debugging, and refining code. For mathematical proofs, AI was used primarily to review and check the authors' arguments for potential errors, inconsistencies, or gaps. AI-assisted manuscript revisions, literature-search suggestions, feedback on research ideas and proofs, and generated or modified code were reviewed and verified by the authors before incorporation into the final work. The authors independently assessed the accuracy and appropriateness of these outputs and retain full responsibility for the methodology, proofs, implementation, analysis, and conclusions presented in the paper.

REPRODUCIBILITY STATEMENT

The code and an accompanying `README` file are submitted with the manuscript. We provide a dedicated reproduction script for each figure and table, allowing the reported results to be reproduced individually. The `README` provides instructions for running the code.

REFERENCES

S. An and X. Huo. A Polyak–Ruppert central limit theorem for SA–Adam with momentum and non-convergent adaptive preconditioning. *arXiv preprint arXiv:2606.17364*, 2026a.

S. An and X. Huo. When does dynamic preconditioning preserve the Polyak–Ruppert CLT? a stabilization threshold. *arXiv preprint arXiv:2604.23498*, 2026b.

P. Billingsley. *Convergence of Probability Measures*. Wiley, 2 edition, 1999.

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

H.-F. Chen. *Stochastic Approximation and Its Applications*. Kluwer Academic Publishers, 2002.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020. doi: 10.1214/18-AOS1801.

D. Davis, D. Drusvyatskiy, and L. Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *Annals of Statistics*, 52:1485–1508, 2024. doi: 10.1214/24-AOS2401.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018.

M. Fernandes, E. Guerre, and E. Horta. Smoothing quantile regressions. *Journal of Business & Economic Statistics*, 39:338–357, 2021. doi: 10.1080/07350015.2019.1660177.

S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12: 461–529, 2018. doi: 10.1214/18-EJS1395.

A. Godichon-Baggioni. Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *Journal of Statistical Planning and Inference*, 203:1–19, 2019. doi: 10.1016/j.jspi.2019.01.001.

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

X. He, X. Pan, K. M. Tan, and W.-X. Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 232:367–388, 2023. doi: 10.1016/j.jeconom.2021.07.010.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964. doi: 10.1214/aoms/1177703732.

P. J. Huber. Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, 1: 799–821, 1973. doi: 10.1214/aos/1176342503.

L. Jiang, A. Roy, K. Balasubramanian, D. Davis, D. Drusvyatskiy, and S. Na. Online covariance estimation in nonsmooth stochastic approximation. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3079–3123, 2025.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

R. Koenker and G. Bassett, Jr. Regression quantiles. *Econometrica*, 46:33–50, 1978.

H. J. Kushner and J. Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal on Control and Optimization*, 31:1045–1062, 1993. doi: 10.1137/0331047.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2 edition, 2003.

S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7381–7389, 2022. doi: 10.1609/aaai.v36i7.20701.

S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast inference for quantile regression with tens of millions of observations. *Journal of Econometrics*, 249:105673, 2025. doi: 10.1016/j.jeconom.2024.105673.

Jinlong Lei and Uday V. Shanbhag. Variance-reduced accelerated first-order methods: Central limit theorems and confidence statements. *Mathematics of Operations Research*, 50(2):1364–1397, 2025. doi: 10.1287/moor.2021.0068. Published online June 4, 2024.

J. Lelong. Almost sure convergence of randomly truncated stochastic algorithms under verifiable conditions. *Statistics & Probability Letters*, 78:2632–2636, 2008. doi: 10.1016/j.spl.2008.02.034.

J. Lelong. Asymptotic normality of randomly truncated stochastic algorithms. *ESAIM: Probability and Statistics*, 17:105–119, 2013. doi: 10.1051/ps/2011110.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2 edition, 1989.

W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak–Ruppert and non-asymptotic concentration. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2947–2997, 2020.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992. doi: 10.1137/0330046.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent Robbins–Monro process. Technical Report 781, School of Operations Research and Industrial Engineering, Cornell University, 1988.

K. Tang, W. Liu, Y. Zhang, and X. Chen. Acceleration of stochastic gradient descent with momentum by averaging: Finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2024.

P. Tüfekci. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140, 2014. doi: 10.1016/j.ijepes.2014.02.027.

P. Tüfekci and H. Kaya. Combined cycle power plant. UCI Machine Learning Repository, 2014. Creative Commons Attribution 4.0 International.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016.

W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118:393–404, 2023. doi: 10.1080/01621459.2021.1933498.

A Simulation details and additional comparisons

A.1 Configuration and implementation

Table 3 gives the 15 configurations. The formal master seed is 161803399. Configuration and replication seeds were fixed before evaluation. Initial parameters and filter states are zero; updates are unprojected.

The frozen pilot grid varied the common decay exponent, a multiplier of the initial stepsize scales, burn-in fraction and horizon multiplier; β and filter shapes were prespecified. Screening used 20 replications and independent confirmation used 50. The criterion minimized median absolute online and oracle coverage errors plus an excess-RMSE penalty, subject to prespecified stability, coverage-improvement, RMSE and exact least-squares bias-reduction requirements. The selected rule sets $\alpha = 0.600$, doubles the initial stepsize scales, uses 30% burn-in and quadruples the initial horizons. These choices were fixed before the 500 evaluation replications.

Recursive Wald intervals use $N_n = n - b_n$ in their standard errors, while batch intervals and full-average self-normalization use n. The score covariance uses retained observations; the online Jacobian uses the full history. A fixed ridge 1.000×10^{-6} is applied as in (5), or directly to the quantile Jacobian. This numerical choice approximates, but is distinct from, the vanishing ridge in the consistency theorems. Numerical failures count as noncoverage, and RMSE is unavailable for a method with failed replications.

The filters and their gain matching are specified in Section 4. All matched filters have identity long-run gain.

Table 3: Configurations for smooth and quantile models. Each row uses $M = 500$, $\alpha = 0.600$ and 30% burn-in. The coefficient β specifies scalar momentum and the second pole.

Design	Configuration	n	d	β	η_0
Smooth	LS, Gaussian, small	8000	5	0.5	0.400
Smooth	LS, t_3, medium	40000	20	0.9	0.360
Smooth	LS, asymmetric, high d	8000	100	0.990	0.160
Smooth	LS, Gaussian, large n	200000	5	0.9	0.400
Smooth	Logit, correct, small	8000	5	0.5	1.600
Smooth	Logit, correct, medium	40000	20	0.9	1.400
Smooth	Logit, misspecified	40000	20	0.990	0.9
Smooth	Logit, misspecified, large n	200000	5	0.9	1.400
Quantile	$\tau = 0.1$, Gaussian	40000	5	0.9	0.360
Quantile	$\tau = 0.5$, t_3, heteroscedastic	40000	5	0.9	0.320
Quantile	$\tau = 0.9$, asymmetric, heteroscedastic	40000	5	0.9	0.320
Quantile	$\tau = 0.5$, Gaussian, small	8000	5	0.9	0.360
Quantile	$\tau = 0.5$, t_3, large n	200000	5	0.9	0.240
Quantile	$\tau = 0.1$, asymmetric	40000	5	0.9	0.280
Quantile	$\tau = 0.9$, t_3	40000	5	0.9	0.280

A.2 Least squares and logistic regression

For the median-centered exponential errors used in the main least-squares comparison, independence of X and ε and $\mathbb{E}X = 0$ imply $\mathbb{E}(X\varepsilon) = 0$. Thus the generating slope solves the estimating equation despite the nonzero error mean.

For each misspecified logistic configuration, the reference computation replaces the population expectation by the average over 150000 independent reference covariates and solves the resulting score equation by Newton iteration. It uses $p_{\mathrm{true}}(X)$ rather than simulated binary responses. The same numerical approximation is used for coverage and RMSE in all 500 replications of that configuration. The reported evaluation does not account for reference-integration error; the population root, rather than its fixed numerical approximation, is the target in the inferential theorem.

Table 4: Filter shape and effective gain in the smooth models. Each numerical entry gives 95% coverage of the first slope/per-coordinate RMSE. The raw-gain method uses the SGD stepsize without momentum normalization. All methods use tail averages and online intervals as in Table 1. A dash marks a raw-gain recursion with at least one numerical failure; success-only RMSE is not reported.

Configuration	SGD	Proposed (momentum)	Proposed (coordinate)	Proposed (two-pole)	Raw gain
LS, Gaussian, small	0.960/0.016	0.960/0.016	0.960/0.016	0.960/0.016	0.960/0.016
LS, t_3, medium	0.950/0.008	0.950/0.008	0.950/0.008	0.950/0.008	–
LS, asymmetric, high d	0.910/0.037	0.964/0.017	0.964/0.019	0.964/0.017	–
LS, Gaussian, large n	0.928/0.003	0.928/0.003	0.928/0.003	0.928/0.003	0.936/0.003
Logit, correct, small	0.948/0.034	0.948/0.034	0.948/0.034	0.948/0.034	0.942/0.035
Logit, correct, medium	0.954/0.016	0.954/0.016	0.954/0.016	0.952/0.016	0.214/0.024
Logit, misspecified	0.960/0.015	0.958/0.015	0.960/0.015	0.958/0.015	0.000/0.126
Logit, misspecified, large n	0.954/0.007	0.954/0.007	0.954/0.007	0.954/0.007	0.954/0.007

A.3 Quantile regression

The quantile data-generating model and the centering $\varepsilon_\tau = U - F_U^{-1}(\tau)$ are specified in Section 4. The configurations labeled heteroscedastic use $\sigma(X) = 0.5 + 0.5|X_2|$; the others use $\sigma(X) = 1$. Gaussian covariates have the AR(1) correlation of the smooth designs. The t_5 covariates apply its Cholesky factor to independent $t_5/\sqrt{5/3}$ variables. The latter are used for median-t_3 heteroscedastic, median-t_3 large-n, and lower-tail asymmetric errors; the remaining four configurations use Gaussian covariates.

Beyond the primary Epanechnikov kernel with $h_t = t^{-0.300}$, selected configurations also evaluate exponents 0.200 and 0.400 and a uniform kernel. The main text specifies the oracle, equal-block and self-normalized implementations.

Table 5 also reports a logistic-convolution prototype minimizing the average of $\tau u + h\log\{1 + \exp(-u/h)\}$, where $u = Y - X^\mathsf{T}\theta$ and $h = \max\{1.000 \times 10^{-4}, \mathrm{sd}(Y)n^{-1/5}\}$. BFGS starts at the batch Koenker–Bassett fit. Both batch intervals use the unsmoothed quantile score and a Gaussian residual-density Jacobian with bandwidth $1.060\,\mathrm{sd}(u)n^{-1/5}$ at their own fitted coefficients. The prototype has no smoothing-bias correction in its interval and is not a fully calibrated implementation of Fernandes et al. (2021); He et al. (2023).

Table 5: Additional quantile comparisons: intercept 95% coverage/per-coordinate RMSE. Recursive intervals use the last 70% of iterates and the Epanechnikov online Jacobian with $\kappa = 0.300$. The smoothed prototype uses the response-scale bandwidth and uncorrected quantile intervals specified above; it is not a fully calibrated implementation of the cited smoothed methods.

Configuration	n	SGD	Proposed (momentum)	Proposed (two-pole)	Batch QR	Smoothed prototype
$\tau = 0.1$, Gaussian	40000	0.956/0.018	0.960/0.018	0.954/0.018	0.938/0.010	0.004/0.020
$\tau = 0.5$, t_3, heteroscedastic	40000	0.966/0.005	0.970/0.005	0.972/0.005	0.966/0.004	0.976/0.004
$\tau = 0.9$, asymmetric, heteroscedastic	40000	0.946/0.044	0.948/0.045	0.950/0.045	0.948/0.016	0.038/0.024
$\tau = 0.5$, Gaussian, small	8000	0.966/0.020	0.968/0.020	0.968/0.020	0.946/0.016	0.984/0.015
$\tau = 0.5$, t_3, large n	200000	0.956/0.002	0.956/0.002	0.954/0.002	0.966/0.002	0.974/0.002
$\tau = 0.1$, asymmetric	40000	0.940/0.002	0.946/0.002	0.946/0.002	0.980/0.002	0.000/0.029
$\tau = 0.9$, t_3	40000	0.946/0.024	0.948/0.024	0.954/0.024	0.950/0.010	0.000/0.030

The prototype undercovers in several tail and asymmetric settings under this response-scale bandwidth. This reflects the specified implementation, not smoothed quantile regression in general. The comparison with batch QR separates this effect from that of filtering.

B POWER-PLANT QUANTILE ILLUSTRATION

B.1 DATA AND STATISTICAL QUESTION

We use the Combined Cycle Power Plant data from the UCI Machine Learning Repository (Tüfekci & Kaya, 2014; Tüfekci, 2014). The dataset has DOI 10.24432/C5002N and a CC BY 4.0 license. Its 9568 sampling units are hourly averages from full-load operation during 2006–2011. There are no reported missing values. Each record gives ambient temperature, exhaust vacuum, ambient pressure, relative humidity, and net electrical output.

The archive provides five shuffled copies of the records but no timestamps. We treat Sheet 1 as a supplied processing order, not a historical stream. Sheets 2–5 measure order sensitivity and are not additional samples. The question is descriptive: how are the four recorded predictors jointly associated with lower, median, and upper linear conditional-quantile projections of net output within the observed full-load regime? The covariance calculation uses working independence.

B.2 PREPROCESSING AND RECURSIVE FITS

We retain all rows and all four predictors. Predictors and output are centered and scaled using Sheet 1 before each method makes one recursive pass through a supplied order. Thus the recursion is one pass after full-sample preprocessing. The intervals under working independence treat the estimated preprocessing constants as fixed; because those constants use all observations, this descriptive convention is not covered directly by the i.i.d. predictable-data theorem. All coefficient estimates and interval endpoints are converted back to MW; slopes are reported per sample standard deviation of a predictor.

We fit linear quantile regressions at $\tau = 0.1, 0.5, 0.9$ with an intercept and all four predictors. The recursive methods are SGD, normalized momentum with $\beta = 0.9$, and the stable two-pole filter, all with $\alpha = 0.600$, $\eta_0 = 0.800$, and a 10% primary burn-in. The primary intervals use the online Epanechnikov Jacobian with $\kappa = 0.300$ and a fixed ridge of 1.000×10^{-5}; full-average self-normalized intervals and batch Koenker–Bassett estimates are retained as comparators. This fixed numerical regularizer differs from the vanishing ridge in Theorem 3.11. The same preprocessing constants and tuning are used for all five supplied orders; stability of these unprojected trajectories is not guaranteed by the projected-recursion result.

B.3 RESULTS

Table 6: Adjusted linear quantile associations in the power-plant data and sensitivity to observation order under working independence. Filtered PR estimates use normalized momentum and intervals with an online kernel Jacobian; slopes are in MW for an increase of one sample standard deviation in the predictor. The second panel gives, for each recursion, the largest coefficient range among temperature, vacuum, and humidity over the five supplied shuffled orders.

τ	Predictor	Filtered PR	95% interval	Batch QR
0.1	Ambient temperature	-14.015	[-14.410, -13.620]	-14.311
0.1	Exhaust vacuum	-3.109	[-3.340, -2.879]	-2.890
0.1	Ambient pressure	0.715	[0.522, 0.908]	0.726
0.1	Relative humidity	-1.849	[-2.027, -1.670]	-2.034
0.5	Ambient temperature	-15.119	[-15.426, -14.811]	-15.182
0.5	Exhaust vacuum	-2.962	[-3.211, -2.712]	-2.939
0.5	Ambient pressure	0.195	[0.062, 0.328]	0.203
0.5	Relative humidity	-2.293	[-2.452, -2.134]	-2.298
0.9	Ambient temperature	-14.431	[-14.751, -14.110]	-14.709
0.9	Exhaust vacuum	-3.576	[-3.847, -3.305]	-3.388
0.9	Ambient pressure	0.165	[-0.020, 0.351]	0.206
0.9	Relative humidity	-2.869	[-3.074, -2.664]	-3.020

Recursion	Largest range across orders
Normalized momentum	0.370
SGD	1.116
Two-pole	0.433

For normalized momentum, the adjusted coefficients of temperature, exhaust vacuum, and relative humidity are negative at all three quantiles, and their intervals under working independence exclude zero. Ambient pressure remains in every fit. Its estimates are 0.715, 0.195, and 0.165 MW per sample standard deviation at $\tau = 0.1, 0.5, 0.9$; the upper-quantile interval includes zero. Table 6 reports all four predictors and the batch estimates.

Across the five supplied orders, the largest slope range among temperature, exhaust vacuum, and relative humidity is 0.370, 0.433, and 1.116 MW for normalized momentum, the two-pole filter, and SGD, respectively. These summarize order sensitivity, not sampling variation. Interpretation is observational and restricted to the fitted model, working independence, and the full-load regime.

C THEORETICAL SUPPLEMENT

C.1 ASYMPTOTIC DISTRIBUTION AND COVARIANCE ESTIMATION

The role of the filter is exposed by an exact change of variables. Define $\mathrm{D} = \mathrm{PB}(\mathrm{I}_r - \mathrm{B})^{-1} \in \mathbb{R}^{d\times r}$ and the transformed error $u_t = e_t - \eta_t \mathrm{D} v_t$, using $\eta_0 = \eta_1$ when defining u_0.

Proposition C.1. *Under Assumption 2.1, the recursion* (1) *satisfies, for every* $t \geq 1$*,*

$$u_t = u_{t-1} - \eta_t \mathrm{K} G(\theta_{t-1}, Z_t) + (\eta_{t-1} - \eta_t)\mathrm{D} v_{t-1}. \tag{17}$$

The identity is algebraic; its proof is given in section C.7.1. In (17), the local mean and target noise are $\mathrm{KA}e$ and $\mathrm{K}\xi_t^\star$, respectively. Their contribution to the averaged influence function is $-(\mathrm{KA})^{-1}\mathrm{K}\xi_t^\star = -\mathrm{A}^{-1}\xi_t^\star$. Thus the gain cancels whenever K is nonsingular. The proof of Theorem 3.4 shows that the remaining terms are negligible under the stated trajectory conditions.

We separate local properties of the estimating equation from properties of the computed trajectory. Write $e_t = \theta_t - \theta_\star$ for the iteration error. We use $\xrightarrow{p}$ and $\xrightarrow{d}$ for convergence in probability and in distribution, respectively, and $o_p(1)$ for a quantity converging to zero in probability.

Assumption 3.1 imposes differentiability of the population score and local identification (van der Vaart, 1998, Chapter 5). Localized bounds use explicit indicators; $\mathfrak{r}(e) = g(\theta_\star + e) - \mathrm{A}e$ retains its definition on $\mathbb{R}^d$. No symmetry or positive definiteness is imposed on A. Assumption 3.2 requires continuity of the centered score in L^2 at $\theta_\star$, but permits discontinuous sample scores.

The rate controls Abel boundary terms, localization permits local score regularity to be used along the path, and the accumulated remainder controls nonlinear mean bias. These are inputs to the inferential theorem; stability of B alone does not imply them. The verification results below give sufficient conditions for scalar momentum.

The full-state bound also constrains scores before localization. Nonsingularity of K forces C to have full column rank, so, with $\mathrm{C}^\dagger = (\mathrm{C}^\mathsf{T}\mathrm{C})^{-1}\mathrm{C}^\mathsf{T}$, the uncorrected recursion gives $G_t = \mathrm{C}^\dagger(v_t - \mathrm{B}v_{t-1})$ and hence $\sup_t \mathbb{E}\|G_t\|^2 < \infty$. Conversely, this score bound and stability of B imply the state bound by lemma C.20. Thus the complete theorem requires more than a target-score second moment. For ordinary SGD, $\mathrm{B} = 0$ and $\mathrm{D} = 0$ in the exact reduction; local mean linearization and trajectory control still govern inference. A sufficient way to check the accumulated remainder is the following.

Proposition C.2. *Suppose that* $\|\mathfrak{r}(e)\| \leq L_g\|e\|^{1+\nu}$ *on* $\mathcal{U}$*, where* $L_g < \infty$ *and* $\nu \in (0, 1]$*. If* (10) *holds and* $\varrho(1+\nu) > 1$*, then* (11) *holds. In particular, the rate* $\mathbb{E}\|e_t\|^2 = O(\eta_t)$ *gives the sufficient condition* $\alpha(1+\nu) > 1$*.*

This proposition is a verification device, not an extra assumption of the main theorem. For a quadratic remainder and the usual mean-square rate, it gives $\alpha > 1/2$. Values $\alpha \leq 1/2$ remain within the general theorem only when localization, the trajectory rate, and the accumulated remainder can be verified under those choices.

Once the asymptotic linear representation has been established, the normal limit for its leading i.i.d. score sum follows from the finite second moment of the target score. For each fixed admissible filter, the limiting covariance is independent of B, C, and P. This is a first-order statement, not a uniform finite-sample comparison: filters close to instability or with poorly conditioned gains can have substantial boundary terms.

Corollary C.3. *Under the assumptions of Theorem 3.4, let* $\widehat{\theta}^Z_{n,b}$ *be a batch* Z*-estimator based on* $Z_{b_n+1}, \ldots, Z_n$ *with the same asymptotic linear expansion as* (12)*. Then* $\sqrt{N_n}(\bar{\theta}_{n,b} - \widehat{\theta}^Z_{n,b}) \xrightarrow{p} 0$*. If* $b_n = o(n)$ *and a full-sample batch estimator* $\widehat{\theta}^Z_n$ *has the corresponding expansion, then* $\sqrt{n}(\bar{\theta}_{n,b} - \widehat{\theta}^Z_n) \xrightarrow{p} 0$*.*

Likewise, let $\bar{\theta}^{(j)}_{n,b}$*,* $j = 1, 2$*, be averages obtained from the same observations and estimating function using fixed filters* $(\mathrm{B}_j, \mathrm{C}_j, \mathrm{P}_j)$*. If both recursions satisfy the assumptions of Theorem 3.4 for the same target and burn-in, then* $\sqrt{N_n}\{\bar{\theta}^{(1)}_{n,b} - \bar{\theta}^{(2)}_{n,b}\} \xrightarrow{p} 0$*.*

The required batch expansion is the regular Z-estimator expansion (van der Vaart, 1998, Chapter 5); section C.7.1 records the normalization for positive-fraction burn-in. Taking $\mathrm{B} = 0$ and $\mathrm{C} = \mathrm{I}_d$

recovers averaged ordinary or fixed-preconditioned stochastic gradient descent (Ruppert, 1988; Polyak & Juditsky, 1992; Kushner & Yang, 1993). Thus admissible fixed filters can change transient and last-iterate behavior, but not the first-order covariance of the average. Any efficiency inherited from a regular batch estimator comes from the estimating equation, not from filtering. Time-varying or adaptive filters require additional stabilization arguments (An & Huo, 2026b;a).

For the estimator, Assumption 3.5 supplies the integrability and score convergence needed along the full trajectory.

The score outer-product law of large numbers follows from the stated L^2 convergence and target second moment; no uniform $(2+\delta)$-moment condition, with $\delta > 0$, is imposed. Batch-means and iterate-only estimators provide alternatives when the Jacobian is difficult to estimate (Godichon-Baggioni, 2019; Zhu et al., 2023; Jiang et al., 2025).

C.2 Verification of the trajectory conditions

We give global and projected routes to trajectory regularity, followed by exact gain matching and other stabilization schemes. Related Lyapunov arguments appear in Kushner & Yin (2003); Borkar (2008).

C.2.1 Sufficient conditions for scalar momentum

The next result supplies the state and rate bounds used above for a familiar filter. It separates a checkable drift condition from the general inferential theorem. For $r = d$, write the scalar-momentum state as $m_t \in \mathbb{R}^d$ and consider

$$m_t = \beta m_{t-1} + cG(\theta_{t-1}, Z_t), \qquad \theta_t = \theta_{t-1} - \eta_t \mathrm{P} m_t, \tag{18}$$

where $0 \le \beta < 1$, $c > 0$, and $\mathrm{P} \in \mathbb{R}^{d\times d}$ is nonsingular. As above, $\mathrm{K} = c\mathrm{P}/(1-\beta)$.

Assumption C.4. *There are constants $\mu > 0$ and $L, \sigma, L_\xi \ge 0$, all finite, such that, for every $e \in \mathbb{R}^d$,*

$$e^{\mathsf{T}} \mathrm{K} g(\theta_\star + e) \ge \mu \|e\|^2, \qquad \|\mathrm{K} g(\theta_\star + e)\| \le L\|e\|,$$
$$\mathbb{E}\|\mathrm{K}\xi(\theta_\star + e, Z_1)\|^2 \le \sigma^2 + L_\xi^2 \|e\|^2.$$

The initial values θ_0 and m_0 are square integrable.

Theorem C.5. *Suppose Assumption C.4 holds. For the recursion* (18) *with $\eta_t = \eta_0 t^{-\alpha}$, $\eta_0 > 0$, and $0 < \alpha < 1$, $\sup_{t\ge 0} \mathbb{E}\|m_t\|^2 < \infty$ and $\mathbb{E}\|\theta_t - \theta_\star\|^2 = O(\eta_t)$. The result holds for every fixed $\eta_0 > 0$. If $\alpha > 1/2$, then $\theta_t \to \theta_\star$ almost surely.*

The rate argument requires a sufficiently small stepsize only after a finite deterministic index, which explains why it does not impose an upper bound on η_0. For $\alpha > 1/2$, almost-sure convergence also gives eventual localization in any neighborhood of the target. A local remainder bound $\|\mathfrak{r}(e)\| \le L_g\|e\|^{1+\nu}$ with $\alpha(1+\nu) > 1$ then verifies all of Assumption 3.3. For smaller α, localization and accumulated bias require their own verification. The global drift bounds are sufficient stability conditions; they are stronger than local identification. For a local route to stability, let $\Pi_{\mathcal{C}}$ denote Euclidean projection onto a closed convex set $\mathcal{C} \subset \mathbb{R}^d$.

Theorem C.6. *Let $\mathcal{C} \subset \mathcal{U}$ be compact and convex, with $\theta_\star$ in its interior, and consider*

$$m_t = \beta m_{t-1} + cG(\theta_{t-1}, Z_t), \qquad \theta_t = \Pi_{\mathcal{C}}\{\theta_{t-1} - \eta_t \mathrm{P} m_t\}. \tag{19}$$

Assume $0 \le \beta < 1$, $c > 0$, P is nonsingular, $\theta_0 \in \mathcal{C}$, and m_0 is square integrable. Suppose every computed score is finite almost surely and the drift and noise inequalities in Assumption C.4 hold for every e such that $\theta_\star + e \in \mathcal{C}$, with $\mathrm{K} = c\mathrm{P}/(1-\beta)$. Suppose also that Assumption 3.1 holds and that $\|\mathfrak{r}(e)\| \le L_r\|e\|^{1+\nu}$ on $\mathcal{C}$, where $L_r < \infty$ and $\nu \in (0, 1]$.

If $\eta_t = \eta_0 t^{-\alpha}$, $\eta_0 > 0$, $0 < \alpha < 1$, and $\alpha(1+\nu) > 1$, then $\sup_{t\ge 0} \mathbb{E}\|m_t\|^2 < \infty$, $\mathbb{E}\|\theta_t - \theta_\star\|^2 = O(\eta_t)$, and $\theta_t \to \theta_\star$ almost surely. The projection changes the unconstrained update only finitely often almost surely. Consequently, Assumption 3.3 holds with $\varrho = \alpha$; under Assumption 3.2 and (2)*, the asymptotic linear representation and normal limit in Theorem 3.4 hold for this projected recursion.*

The restriction $\alpha(1+\nu) > 1$ both makes the accumulated remainder negligible and implies $\alpha > 1/2$, which gives square-summable stepsizes for the almost-sure localization argument. It is not needed for the mean-square rate or state bound alone. The theorem provides a complete trajectory verification for deterministic projection.

The normalization of scalar momentum affects the effective gain but is not essential for the limiting covariance. In fact, matching the gain gives an exact, rather than asymptotic, correspondence.

Proposition C.7. *Let superscripts N and U denote the recursions*

$$m_t^N = \beta m_{t-1}^N + (1-\beta)G(\theta_{t-1}^N, Z_t), \quad \theta_t^N = \theta_{t-1}^N - \eta_t^N \mathrm{P} m_t^N,$$
$$m_t^U = \beta m_{t-1}^U + G(\theta_{t-1}^U, Z_t), \qquad \theta_t^U = \theta_{t-1}^U - \eta_t^U \mathrm{P} m_t^U.$$

Suppose the two recursions use the same observations, score map, preconditioner, and parameter initialization. If $m_0^N = (1-\beta)m_0^U$ and $\eta_t^U = (1-\beta)\eta_t^N$ for every $t \geq 1$, then $m_t^N = (1-\beta)m_t^U$ and $\theta_t^N = \theta_t^U$ for every $t \geq 0$. Zero momentum initialization satisfies the required initial relation.

Thus matched implementations hold $\eta_t \mathrm{K}$ fixed. Using the same raw stepsize for the two normalizations instead compares recursions with different effective drifts.

For reset or expanding-truncation schemes, a separate argument must show that the corrections stop almost surely before proposition C.26 can transfer the inferential limit (Chen, 2002; Lelong, 2008; 2013). Neither that transfer result nor the global mean-square rate is a substitute for scheme-specific stability verification.

C.3 Functional limits and self-normalized inference

A terminal central limit theorem does not by itself justify inference based on the entire solution path. For this subsection, use the full average $\bar{\theta}_n = n^{-1}\sum_{t=1}^n \theta_t$ and define $\mathbb{U}_n(s) = n^{-1/2}\sum_{t=1}^{\lfloor ns \rfloor} e_t$ for $s \in [0,1]$, with an empty sum interpreted as zero. Here $\lfloor x \rfloor$ is the greatest integer not exceeding x. Let $\mathbb{H} = L^2([0,1];\mathbb{R}^d)$ be the Hilbert space of square-integrable vector-valued functions, with $\|x\|_{\mathbb{H}}^2 = \int_0^1 \|x(s)\|^2\, ds$. Write $\mathrm{S}^{1/2}$ for the symmetric positive-semidefinite square root of S, set $\Lambda = -\mathrm{A}^{-1}\mathrm{S}^{1/2}$, and let W_d be standard d-dimensional Brownian motion. Thus $\Lambda\Lambda^{\mathsf{T}} = \Sigma$.

No additional moment or smoothness assumption is needed beyond those of the terminal limit. Joint convergence includes the endpoint explicitly because evaluation at $s = 1$ is not a continuous functional on L^2. The topology is weaker than uniform path convergence but is sufficient for the integrated self-normalizer below. A stronger path-space limit requires additional control of the largest transformed-state boundary term.

Proposition C.8. *In addition to the assumptions of Theorem 3.7, suppose*

$$\frac{1}{\sqrt{n}} \max_{1 \leq k \leq n} \frac{\|u_k\|}{\eta_k} \xrightarrow{p} 0, \tag{20}$$

where $u_k = e_k - \eta_k \mathrm{D} v_k$ is the transformed error in proposition C.1. Then $\mathbb{U}_n \xrightarrow{d} \Lambda W_d$ in $D([0,1];\mathbb{R}^d)$, the space of right-continuous functions with left limits, under the Skorokhod topology. The polygonal interpolation also converges in $C([0,1];\mathbb{R}^d)$, the space of continuous functions, under the uniform norm.

A sufficient condition for (20) *is the existence of $p \geq 2$, $q_u > 2\alpha - 1$, and $C_u < \infty$ such that $\mathbb{E}\|u_t\|^p \leq C_u t^{-pq_u/2}$ and $p(q_u + 1 - 2\alpha) > 2$.*

Condition (20) is needed only for the path-space upgrade, not for the self-normalized inference that follows.

For an integer $q \geq 1$, let W_q denote standard q-dimensional Brownian motion and write

$$\mathrm{J}_q = \int_0^1 \{W_q(s) - sW_q(1)\}\{W_q(s) - sW_q(1)\}^{\mathsf{T}}\, ds.$$

The normalizer in (8) and its Brownian reference laws are established constructions for unfiltered procedures: Lee et al. (2022) use random scaling for averaged stochastic gradient descent, and Lee

et al. (2025) develop solution-path inference for stochastic-subgradient quantile regression. The contribution here is the transfer to the fixed stable filters and discontinuous-score estimating equations covered by Theorem 3.7. The reference laws in (13) and (14) involve no unknown model parameters and can be simulated once. This avoids Jacobian and score-covariance estimation, bandwidth selection, batch length selection, and resampling, at the cost of a non-Gaussian reference law and potentially wider intervals. The coverage statement is fixed-horizon asymptotic inference; it does not establish optional-stopping or anytime-valid coverage.

The normalizer can be computed without storing the path. For a selected contrast matrix R_0, let $T_k = \mathrm{R}_0 \sum_{t=1}^{k} \theta_t \in \mathbb{R}^q$, $\mathrm{M}_n^{(2)} = \sum_{k=1}^{n} T_k T_k^{\mathsf{T}} \in \mathbb{R}^{q\times q}$, and $M_n^{(1)} = \sum_{k=1}^{n} kT_k \in \mathbb{R}^q$. Expanding (8) gives

$$\mathrm{R}_0 \widehat{\mathrm{V}}_n \mathrm{R}_0^{\mathsf{T}} = \frac{1}{n^2}\left\{ \mathrm{M}_n^{(2)} - \frac{M_n^{(1)} T_n^{\mathsf{T}} + T_n (M_n^{(1)})^{\mathsf{T}}}{n} + \frac{\sum_{k=1}^{n} k^2}{n^2} T_n T_n^{\mathsf{T}} \right\}.$$

After forming $\mathrm{R}_0\theta_t$, the required summaries use $O(q^2)$ memory and arithmetic per observation; dense contrast multiplication costs $O(qd)$. A small number of prespecified contrasts therefore avoids storing the normalizer in the full parameter dimension.

C.4 Model specializations

The preceding results apply to both smooth and nonsmooth estimating functions. The specializations below identify the population matrices and the score regularity needed for inference. Unless a projected trajectory result is invoked explicitly, they remain conditional on Assumption 3.3 for the chosen filter.

For smooth empirical-risk minimization, let $\ell(\theta, Z) \in \mathbb{R}$ be a differentiable sample loss, where Z has the same distribution as Z_1, and define the population risk $R(\theta) = \mathbb{E}\{\ell(\theta, Z)\}$. Use the score $G(\theta, Z) = \nabla_\theta \ell(\theta, Z)$.

Corollary C.9. *Suppose* $\mathbb{E}\{\nabla_\theta \ell(\theta, Z)\} = \nabla R(\theta)$ *near* $\theta_\star$, $\nabla R(\theta_\star) = 0$, *and* $\mathrm{H} = \nabla^2 R(\theta_\star)$ *is nonsingular, with* $\nabla R(\theta_\star + e) = \mathrm{H}e + o(\|e\|)$. *Assume* $\mathbb{E}\|\nabla_\theta \ell(\theta_\star, Z)\|^2 < \infty$, *the centered stochastic gradient satisfies Assumption 3.2, and Assumptions 2.1 and 3.3 and* (2) *hold with* $\mathrm{A} = \mathrm{H}$. *Then* $\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_\star) \xrightarrow{d} \mathcal{N}_d(0, \mathrm{H}^{-1}\mathrm{V}_\ell \mathrm{H}^{-\mathsf{T}})$, *where* $\mathrm{V}_\ell = \mathrm{Var}\{\nabla_\theta \ell(\theta_\star, Z)\}$.

A deterministic differentiable penalty may be included in R. It contributes to the mean derivative H, but not to the stochastic variance V_ℓ unless the penalty itself is randomized.

For the regression models, write $Z = (X, Y)$ and $Z_t = (X_t, Y_t)$, where $X \in \mathbb{R}^d$ is the regressor vector and $Y \in \mathbb{R}$ is the response. In least squares, suppose $\mathbb{E}\|XY\| < \infty$, let $\mathrm{Q} = \mathbb{E}(XX^{\mathsf{T}})$ be finite and positive definite, and define $\theta_{\mathrm{LS}} = \mathrm{Q}^{-1}\mathbb{E}(XY)$ and $\varepsilon = Y - X^{\mathsf{T}}\theta_{\mathrm{LS}}$. The score is $G(\theta; X, Y) = X(X^{\mathsf{T}}\theta - Y)$. This target is defined by the population estimating equation and does not require a correctly specified conditional mean model.

Corollary C.10. *Suppose* $\mathbb{E}(\varepsilon^2 \|X\|^2) < \infty$ *and, for some* $p \in (1, 2]$,

$$\mathbb{E}\|XX^{\mathsf{T}} - \mathrm{Q}\|_{\mathrm{F}}^{p} < \infty, \qquad p(1 + \varrho) > 2, \tag{21}$$

where ϱ *is the exponent in* (10). *Suppose Assumptions 2.1 and 3.3 and* (2) *hold with* $\theta_\star = \theta_{\mathrm{LS}}$, $\mathcal{U} = \mathbb{R}^d$, $\mathrm{A} = \mathrm{Q}$, *and* $\mathfrak{r} \equiv 0$. *Then* $\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_{\mathrm{LS}}) \xrightarrow{d} \mathcal{N}_d(0, \Sigma_{\mathrm{LS}})$, *where* $\Sigma_{\mathrm{LS}} = \mathrm{Q}^{-1}\mathbb{E}(\varepsilon^2 XX^{\mathsf{T}})\mathrm{Q}^{-1}$. *If* $\mathbb{E}(\varepsilon^2 \mid X) = \sigma^2$ *for a constant* σ^2, *this covariance reduces to* $\sigma^2\mathrm{Q}^{-1}$.

The least-squares mean equation is exactly linear, so there is no accumulated nonlinear remainder. Condition (21) controls score variation without itself imposing a fourth regressor moment; when $\varrho = \alpha$, any $p > 2/(1 + \alpha)$ suffices. However, verifying the uniform filter-state bound for an unprojected least-squares recursion may still require a fourth regressor moment. The corollary therefore allows heavy-tailed errors with finite target-score variance, but is not an end-to-end fourth-moment-free stability result for the regressors. For scalar momentum, Theorem C.5 applies when the symmetric part of KQ is positive definite and the stated noise-growth bound holds.

For logistic regression, assume $Y \in \{0, 1\}$. Define $\pi(s) = \{1 + \exp(-s)\}^{-1}$ for $s \in \mathbb{R}$, write $\pi_\theta(X) = \pi(X^{\mathsf{T}}\theta)$, and use the estimating function $G(\theta; X, Y) = X\{\pi_\theta(X) - Y\}$. Let $\theta_{\log}$ be a

root of the population score, and write $\pi_\star = \pi_{\theta_{\log}}(X)$. The root may be a pseudo-true parameter under misspecification.

Corollary C.11. *Suppose* $\mathbb{E}\|X\|^2 < \infty$ *and* $\mathrm{A}_{\log} = \mathbb{E}\{\pi_\star(1-\pi_\star)XX^\mathsf{T}\}$ *is positive definite. If Assumptions 2.1 and 3.3 and* (2) *hold for the logistic target, then* $\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_{\log}) \xrightarrow{d} \mathcal{N}_d(0, \Sigma_{\log})$, *where* $\mathrm{S}_{\log} = \mathbb{E}[XX^\mathsf{T}(Y-\pi_\star)^2]$ *and* $\Sigma_{\log} = \mathrm{A}_{\log}^{-1}\mathrm{S}_{\log}\mathrm{A}_{\log}^{-1}$. *Under the correctly specified model* $\mathbb{P}(Y=1 \mid X) = \pi_\star$, *one has* $\mathrm{S}_{\log} = \mathrm{A}_{\log}$ *and* $\Sigma_{\log} = \mathrm{A}_{\log}^{-1}$.

A second regressor moment is enough for local differentiability of the mean and local quadratic-mean score continuity. The stronger condition $\mathbb{E}\|X\|^{2+\nu} < \infty$, with $\nu \in (0,1]$, gives a mean remainder of order $\|e\|^{1+\nu}$ and verifies (11) when $\varrho(1+\nu) > 1$. Because the logistic drift is not globally strongly monotone, projection provides a useful route to the required trajectory conditions (McCullagh & Nelder, 1989).

Corollary C.12. *Suppose* $\theta_{\log}$ *is a root of the logistic population score,* $\mathrm{A}_{\log}$ *is positive definite, and* $\mathbb{E}\|X\|^{2+\nu} < \infty$ *for some* $\nu \in (0,1]$. *Choose* $\mathrm{P} = p_0\mathrm{I}_d$ *with* $p_0 > 0$, *and let* $\mathcal{C}$ *be a deterministic compact convex set containing* $\theta_{\log}$ *in its interior. Run* (19) *with* $0 \le \beta < 1$, $c > 0$, $\theta_0 \in \mathcal{C}$, *square-integrable* m_0, *and* $\eta_t = \eta_0 t^{-\alpha}$, *where* $\eta_0 > 0$, $0 < \alpha < 1$, *and* $\alpha(1+\nu) > 1$. *Then projection is active only finitely often almost surely, Assumption 3.3 holds with* $\varrho = \alpha$, *and, under* (2), $\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_{\log}) \xrightarrow{d} \mathcal{N}_d(0, \Sigma_{\log})$.

The projection set must be specified before the run and known to contain the target, for example through an a priori parameter bound. A set based on a data-dependent pilot needs its own coverage and stability argument. The projected result does not establish stability of the unprojected algorithm from the same initial values.

For Huber regression, fix the clipping constant $c_H > 0$. Define $\psi_{c_H}(u) = \max\{-c_H, \min(c_H, u)\}$. Let θ_H solve $\mathbb{E}[X\psi_{c_H}(X^\mathsf{T}\theta_H - Y)] = 0$, set $\varepsilon_H = Y - X^\mathsf{T}\theta_H$, and, for $s \in \mathbb{R}$, write $p_X(s) = \mathbb{P}(|s - \varepsilon_H| < c_H \mid X)$. The estimating function is $G_H(\theta; X, Y) = X\psi_{c_H}(X^\mathsf{T}\theta - Y)$.

Assumption C.13. *The target* θ_H *is a root of the Huber estimating equation,* $\mathbb{E}\|X\|^2 < \infty$, *and*

$$\left\| \mathbb{E}\left[X \int_0^{X^\mathsf{T}e} \{p_X(s) - p_X(0)\}\, ds \right] \right\| = o(\|e\|) \qquad (e \to 0). \tag{22}$$

The matrix $\mathrm{A}_H = \mathbb{E}\{p_X(0)XX^\mathsf{T}\}$ *is finite and positive definite.*

Corollary C.14. *Suppose Assumptions 2.1, 3.3 and C.13 and* (2) *hold for the Huber target. Then* $\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_H) \xrightarrow{d} \mathcal{N}_d(0, \Sigma_H)$, *where* $\mathrm{S}_H = \mathbb{E}\{XX^\mathsf{T}\psi_{c_H}(\varepsilon_H)^2\}$ *and* $\Sigma_H = \mathrm{A}_H^{-1}\mathrm{S}_H\mathrm{A}_H^{-1}$.

The clipped score is continuous but not differentiable at its clipping points. Condition (22) instead requires differentiability of the population score. Let $\nu_H \in (0,1]$ and let $L_H(X)$ be a nonnegative envelope satisfying $\mathbb{E}\{L_H(X)\|X\|^{2+\nu_H}\} < \infty$. One sufficient condition is $|p_X(s) - p_X(0)| \le L_H(X)|s|^{\nu_H}$. This yields a mean remainder of order $\|e\|^{1+\nu_H}$ and verifies the accumulated-remainder condition when $\varrho(1+\nu_H) > 1$. It is a sufficient regularity condition, not part of the estimator definition (Huber, 1964; 1973).

These three regression models admit direct online Jacobian estimates.

Proposition C.15. *In least squares,* $\mathbb{E}\|X\|^2 < \infty$ *implies* $\widehat{\mathrm{A}}_{\mathrm{LS},n} = n^{-1}\sum_{t=1}^n X_t X_t^\mathsf{T} \xrightarrow{p} \mathrm{Q}$. *For logistic and Huber regression, suppose the relevant trajectory converges in probability to its target. In logistic regression, the same moment condition gives*

$$\widehat{\mathrm{A}}_{\log,n} = \frac{1}{n}\sum_{t=1}^n \pi_{\theta_{t-1}}(X_t)\{1 - \pi_{\theta_{t-1}}(X_t)\}X_t X_t^\mathsf{T} \xrightarrow{P} \mathrm{A}_{\log}.$$

In Huber regression, if $\mathbb{E}\|X\|^2 < \infty$ *and* $\mathbb{E}[\|X\|^2 \mathbf{1}\{|\varepsilon_H| = c_H\}] = 0$, *then*

$$\widehat{\mathrm{A}}_{H,n} = \frac{1}{n}\sum_{t=1}^n X_t X_t^\mathsf{T} \mathbf{1}\{|Y_t - X_t^\mathsf{T}\theta_{t-1}| < c_H\} \xrightarrow{P} \mathrm{A}_H.$$

The weighted boundary condition for Huber regression already follows from Assumption C.13; its role for the standalone Jacobian estimate is made explicit here. The squared logistic and Huber scores are

uniformly integrable, and the score convergence required in Assumption 3.5 holds. For least squares, Assumption 3.5 requires a separate uniform-integrability or moment argument for $X_t(X_t^\mathsf{T}\theta_{t-1} - Y_t)$. Consistency in each case also requires a nonsingular limiting Jacobian, a vanishing positive ridge, and the remaining conditions of Theorem 3.6.

We now consider a discontinuous estimating function. Fix a quantile level $\tau \in (0,1)$ and define $G_\tau(\theta; X, Y) = X\{\mathbf{1}(Y \le X^\mathsf{T}\theta) - \tau\}$. Let $g_\tau(\theta) = \mathbb{E}\{G_\tau(\theta; X, Y)\}$ and let θ_τ be a root of g_τ. Write $F(y \mid X) = \mathbb{P}(Y \le y \mid X)$ and $q_\tau(X) = X^\mathsf{T}\theta_\tau$. This target definition permits misspecification. Solution-path inference is available for unfiltered stochastic-subgradient quantile regression (Lee et al., 2025); the results here also accommodate a fixed stable filter and direct estimation.

Proposition C.16. *Under Assumption 3.9, let $\mathfrak{r}_\tau(e) = g_\tau(\theta_\tau + e) - \mathrm{A}_\tau e$ be the mean remainder specialized to the quantile score. Then $g_\tau(\theta_\tau + e) = \mathrm{A}_\tau e + \mathfrak{r}_\tau(e)$ with $\|\mathfrak{r}_\tau(e)\| = o(\|e\|)$, and $\mathbb{E}\|G_\tau(\theta_\tau + e; Z) - G_\tau(\theta_\tau; Z)\|^2 \to 0$ as $e \to 0$. In particular, the centered quantile score satisfies Assumption 3.2.*

A sufficient condition for (15) is the existence of a conditional density $f(\cdot \mid X)$ with $f_\tau(X) = f(q_\tau(X) \mid X)$ and $|f(q_\tau(X) + s \mid X) - f(q_\tau(X) \mid X)| \le L_f(X)|s|^{\nu_\tau}$, where $\nu_\tau \in (0,1]$ and $L_f(X)$ is a nonnegative envelope satisfying $\mathbb{E}\{L_f(X)\|X\|^{2+\nu_\tau}\} < \infty$. The resulting mean remainder is $O(\|e\|^{1+\nu_\tau})$ (Koenker, 2005). A full conditional density is not otherwise required for the central limit theorem: the regressor second moment and no-atom condition give the needed quadratic-mean score continuity.

Theorem C.17. *Suppose Assumptions 2.1, 3.3 and 3.9 and (2) hold, with $\theta_\star = \theta_\tau$, $\mathrm{A} = \mathrm{A}_\tau$, and mean remainder $\mathfrak{r}_\tau$. Set $\mathrm{S}_\tau = \mathbb{E}[XX^\mathsf{T}\{\mathbf{1}(Y \le X^\mathsf{T}\theta_\tau) - \tau\}^2]$, the target-score covariance. Then $\sqrt{N_n}(\bar\theta_{n,b} - \theta_\tau) \xrightarrow{d} \mathcal{N}_d(0, \Sigma_\tau)$, where $\Sigma_\tau = \mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}$. Under the correctly specified conditional quantile model $\mathbb{P}(Y \le X^\mathsf{T}\theta_\tau \mid X) = \tau$ almost surely, $\mathrm{S}_\tau = \tau(1-\tau)\mathbb{E}(XX^\mathsf{T})$. If $b_n = o(n)$ and the Koenker–Bassett estimator has its usual Bahadur expansion under the additional empirical-process conditions for batch quantile regression, the filtered average is first-order equivalent to that estimator.*

The final equivalence is conditional on the batch expansion; it does not follow from the trajectory assumptions alone (Koenker & Bassett, 1978; Koenker, 2005).

For online estimation of A_τ, let $\mathcal{K} : \mathbb{R} \to [0, \infty)$ be a bounded, compactly supported kernel satisfying $\int \mathcal{K}(u)\,du = 1$ and $\int \mathcal{K}(u)^p\,du < \infty$ for every $p \in [1, 2]$. Choose a bandwidth $h_t = h_0 t^{-\kappa}$ with $h_0 > 0$ and $0 < \kappa < 1$.

The kernel estimator in (16) requires additional density regularity, distinct from the conditions needed for Theorem C.17.

Since A_τ is positive definite and the kernel estimate is positive semidefinite, this specialization also permits a ridge applied directly to the Jacobian. Set $\widetilde{\mathrm{A}}_{\tau,n} = \widehat{\mathrm{A}}_{\tau,n} + \lambda_n \mathrm{I}_d$, with the positive vanishing ridge defined above, and let $\widehat{\mathrm{S}}_{\tau,n,b} = N_n^{-1}\sum_{t=b_n+1}^{n} G_\tau(\theta_{t-1}; X_t, Y_t)G_\tau(\theta_{t-1}; X_t, Y_t)^\mathsf{T}$.

The trajectory rate in Assumption 3.3 is one sufficient way to verify the displacement-moment condition. The theorem proves covariance consistency separately from the averaged-iterate limit; Wald inference also requires Theorem C.17. In the kernel proof, integrating over the fresh response gives a bias bounded by $C\|\theta_{t-1} - \theta_\tau\|^{\nu_A} + Ch_t^{\nu_A}$. Consequently, no condition comparing the parameter error to h_t is needed under the stated global density envelope.

The density envelope and weighted moments in Assumption 3.10 are used for the kernel estimator, not for the averaged-iterate limit in Theorem C.17. The Jacobian result allows any $p > 1$ within the stated range; a weighted fourth moment is needed only when the optional choice $p = 2$ is made.

C.5 A VERIFIABLE NONSMOOTH MODEL

An explicit projected quantile model is verified in Example C.18.

C.6 ONE-PASS IMPLEMENTATION

For an online run, choose a fixed integer burn-in $b \geq 0$ before processing observations. The sequence b_n in (2) describes the more general theoretical reporting window. To compute self-normalization for R_0, maintain $T_t = \sum_{j=1}^{t} \mathrm{R}_0 \theta_j \in \mathbb{R}^q$, $M_t^{(1)} = \sum_{k=1}^{t} kT_k \in \mathbb{R}^q$, and $\mathrm{M}_t^{(2)} = \sum_{k=1}^{t} T_k T_k^{\mathsf{T}} \in \mathbb{R}^{q \times q}$. Expanding (8) gives

$$\mathrm{R}_0 \widehat{\mathrm{V}}_t \mathrm{R}_0^{\mathsf{T}} = \frac{1}{t^2} \Big\{ \mathrm{M}_t^{(2)} - \frac{M_t^{(1)} T_t^{\mathsf{T}} + T_t (M_t^{(1)})^{\mathsf{T}}}{t} + \frac{(t+1)(2t+1)}{6t} T_t T_t^{\mathsf{T}} \Big\}. \tag{23}$$

Equation (23) requires only the current accumulators. For the scalar interval, set $q = 1$ and $\mathrm{R}_0 = a^{\mathsf{T}}$. Algorithm 1 combines the parameter update, estimation, and self-normalization. The Jacobian contributions and scores are evaluated at the parameter value before the current update.

The implementation stores the current parameter, filter state, running mean, and two $d \times d$ matrices. Apart from score and Jacobian evaluation and matrix factorizations at reporting times, dense updates require $O(d^2 + r^2)$ memory and $O(d^2 + r^2 + dr)$ arithmetic per observation. Self-normalization for q prespecified contrasts adds $O(q^2)$ storage and $O(qd + q^2)$ arithmetic. Running only self-normalized inference removes the Jacobian and score-covariance accumulators. Neither procedure stores past observations or iterates.

C.7 PROOFS AND CONSTRUCTIVE VERIFICATION

Proof of proposition C.7. If the identities hold at time $t-1$, the two score evaluations agree. The momentum equations then give $m_t^N = (1-\beta) m_t^U$, and the matched stepsizes make the parameter increments equal. The result follows by induction from the stated initialization. □

Example C.18 (A projected quantile model)**.** *Let $X = (1, U_1, \ldots, U_{d-1})^{\mathsf{T}}/\sqrt{d}$, where the U_j are independent Rademacher variables, so that $\|X\| = 1$ and $\mathrm{Q} = \mathrm{I}_d/d$. Suppose the parameter is known a priori to satisfy $\|\theta_\tau\| \leq 1$, and set $Y = X^{\mathsf{T}}\theta_\tau + \varepsilon$, where ε is independent of X and satisfies $\varepsilon = T - \log\{\tau/(1-\tau)\}$, with T standard logistic. Then $F_\varepsilon(0) = \tau$, $f_\varepsilon(0) = \tau(1-\tau)$, and*

$$\mathrm{A}_\tau = \mathrm{S}_\tau = \frac{\tau(1-\tau)}{d} \mathrm{I}_d, \qquad \Sigma_\tau = \frac{d}{\tau(1-\tau)} \mathrm{I}_d.$$

Take the known feasible set $\mathcal{C} = \{\theta : \|\theta\| \leq 2\}$, $\theta_0 = m_0 = 0$, fixed $0 \leq \beta < 1$, $c = 1 - \beta$, and $\mathrm{P} = \mathrm{I}_d$. For $\theta \in \mathcal{C}$, $|X^{\mathsf{T}}(\theta - \theta_\tau)| \leq 3$. With $m_\tau = \min_{|s| \leq 3} f_\varepsilon(s) > 0$,

$$e^{\mathsf{T}} g_\tau(\theta_\tau + e) \geq (m_\tau/d)\|e\|^2, \qquad \|g_\tau(\theta_\tau + e)\| \leq \tfrac{1}{4}\|e\|.$$

The score is bounded by one, the mean remainder is $O(\|e\|^2)$, and the density is globally bounded and Lipschitz. Thus the projected trajectory theorem, quantile CLT, and kernel-Jacobian theorem all apply, for example with $\alpha = 3/4$, $h_t = h_0 t^{-1/3}$, $\lambda_n = n^{-1/4}$, $h_0, \eta_0 > 0$, and the Epanechnikov kernel $\mathcal{K}(u) = \frac{3}{4}(1-u^2)\mathbf{1}\{|u| \leq 1\}$. The joint integrated limit and self-normalized inference follow from proposition C.26.

C.7.1 PROOFS FOR THE FILTER REDUCTION AND ASYMPTOTIC LINEARITY

Preliminary matrix and summation lemmas

Lemma C.19. *If $\rho(\mathrm{B}) < 1$, then there are constants $M_B < \infty$ and $q_B \in (0,1)$ such that*

$$\left\|\mathrm{B}^j\right\| \leq M_B q_B^j, \qquad j \geq 0. \tag{24}$$

Proof. Choose q_B strictly between $\rho(\mathrm{B})$ and 1. By the spectral-radius formula, $\lim_{j\to\infty} \left\|\mathrm{B}^j\right\|^{1/j} = \rho(\mathrm{B}) < q_B$. Hence there is j_0 such that $\left\|\mathrm{B}^j\right\| \leq q_B^j$ for $j \geq j_0$. Taking

$$M_B = \max\left\{1, \max_{0 \leq j < j_0} \left\|\mathrm{B}^j\right\| q_B^{-j}\right\}$$

gives (24) for every j. □

Algorithm 1 Polyak-Ruppert inference with a stable filter

Require: Initial values θ_0, v_0; fixed $\mathrm{B}, \mathrm{C}, \mathrm{P}$; stepsizes η_t; a fixed integer burn-in $b \geq 0$; reporting ridges $\lambda_t \downarrow 0$; a problem-specific Jacobian update rule; a nonzero contrast $a \in \mathbb{R}^d$ and level $\gamma \in (0,1)$; prespecified deterministic reporting times.

Require: Select no self-normalization, scalar self-normalization for a, or vector self-normalization. In scalar mode set $q = 1$ and $\mathrm{R}_0 = a^\mathsf{T}$. In vector mode supply a prespecified contrast matrix $\mathrm{R}_0 \in \mathbb{R}^{q \times d}$ with $1 \leq q \leq d$ and $\mathrm{rank}(\mathrm{R}_0) = q$, and a null value $r_0 \in \mathbb{R}^q$.

Initialize retained accumulators $\bar{\theta}_{b,b} = 0$, $\widehat{\mathrm{S}}_{b,b} = 0$, and the Jacobian updater.
if full-path self-normalization is requested **then**
    Set $T_0 = 0_q$, $M_0^{(1)} = 0_q$, $\mathrm{M}_0^{(2)} = 0_{q \times q}$.
end if
for $t = 1, 2, \ldots$ **do**
    Observe Z_t; save $\theta^- \leftarrow \theta_{t-1}$; evaluate $G_t \leftarrow G(\theta^-, Z_t)$.
    $v_t \leftarrow \mathrm{B} v_{t-1} + \mathrm{C} G_t$; $\theta_t \leftarrow \theta^- - \eta_t \mathrm{P} v_t$.
    if $t > b$ **then**

$$\bar{\theta}_{t,b} \leftarrow \bar{\theta}_{t-1,b} + (t-b)^{-1}(\theta_t - \bar{\theta}_{t-1,b}),$$

$$\widehat{\mathrm{S}}_{t,b} \leftarrow \widehat{\mathrm{S}}_{t-1,b} + (t-b)^{-1}(G_t G_t^\mathsf{T} - \widehat{\mathrm{S}}_{t-1,b}).$$

    end if
    Update $\widehat{\mathrm{A}}_t$ using the chosen rule evaluated at (θ^-, Z_t), retaining its prescribed full-sample history.
    if full-path self-normalization is requested **then**

$$T_t \leftarrow T_{t-1} + \mathrm{R}_0 \theta_t,$$

$$M_t^{(1)} \leftarrow M_{t-1}^{(1)} + t T_t,$$

$$\mathrm{M}_t^{(2)} \leftarrow \mathrm{M}_{t-1}^{(2)} + T_t T_t^\mathsf{T}.$$

    end if
    if $t > b$ and t is a reporting time **then**
        Form R_t and $\widehat{\Sigma}_{t,b}$ by (5) when its inputs are finite.
        Report (6) with $N_t = t - b$ if its inputs are finite and $a^\mathsf{T} \widehat{\Sigma}_{t,b} a > 0$; otherwise record the Wald report as unavailable.
        if full-path self-normalization is requested **then**

$$\mathrm{R}_0 \widehat{\mathrm{V}}_t \mathrm{R}_0^\mathsf{T} \leftarrow t^{-2}\big\{\mathrm{M}_t^{(2)} - t^{-1}[M_t^{(1)} T_t^\mathsf{T} + T_t (M_t^{(1)})^\mathsf{T}] + \frac{(t+1)(2t+1)}{6t} T_t T_t^\mathsf{T}\big\}.$$

            if $\mathrm{R}_0 \widehat{\mathrm{V}}_t \mathrm{R}_0^\mathsf{T}$ is finite and positive definite **then**
                In scalar mode, report (9). In vector mode, report the test based on (14) for $\mathrm{R}_0 \theta_\star = r_0$. Use the full average contrast T_t/t, not the retained average.
            else
                Record the self-normalized report as unavailable.
            end if
        end if
    end if
end for

Lemma C.20. *Suppose* $\rho(\mathrm{B}) < 1$, $\mathbb{E}\,\|v_0\|^2 < \infty$, *and* $\sup_t \mathbb{E}\,\|G_t\|^2 < \infty$. *Then*

$$\sup_{t \geq 0} \mathbb{E}\,\|v_t\|^2 < \infty.$$

Proof. Iterating the filter recursion gives

$$v_t = \mathrm{B}^t v_0 + \sum_{s=1}^{t} \mathrm{B}^{t-s} \mathrm{C} G_s.$$

By Minkowski's inequality in L^2, lemma C.19, and the uniform score bound,

$$\{\mathbb{E}\,\|v_t\|^2\}^{1/2} \le \left\|\mathrm{B}^t\right\| \{\mathbb{E}\,\|v_0\|^2\}^{1/2} + \sum_{s=1}^{t} \left\|\mathrm{B}^{t-s}\mathrm{C}\right\| \{\mathbb{E}\,\|G_s\|^2\}^{1/2}$$

$$\le M_B\{\mathbb{E}\,\|v_0\|^2\}^{1/2} + M_B\,\|\mathrm{C}\| \sup_s\{\mathbb{E}\,\|G_s\|^2\}^{1/2} \sum_{j=0}^{\infty} q_B^j.$$

The right-hand side is finite and independent of t. Squaring proves the result. □

Lemma C.21. *For $\eta_t = \eta_0 t^{-\alpha}$ with $0 < \alpha < 1$, there is $C_\eta < \infty$ such that, for $t \ge 2$,*

$$0 \le \frac{\eta_{t-1} - \eta_t}{\eta_t} \le \frac{C_\eta}{t}.$$

Proof. The ratio equals

$$\frac{\eta_{t-1}}{\eta_t} - 1 = \left(\frac{t}{t-1}\right)^{\alpha} - 1.$$

Apply the mean value theorem to $x \mapsto x^\alpha$ on $[1, t/(t-1)]$. Its derivative is bounded on $[1, 2]$, and $t/(t-1) - 1 = (t-1)^{-1} \le 2/t$ for $t \ge 2$. □

Lemma C.22. *Let $0 \le b_n < n$ satisfy* (2)*. Suppose u_t is finite almost surely for every t and, for some $q > 2\alpha - 1$,*

$$\mathbb{E}\,\|u_t\|^2 \le Ct^{-q} \tag{25}$$

for all sufficiently large t, where $\eta_t = \eta_0 t^{-\alpha}$ and $0 < \alpha < 1$. Then

$$\frac{1}{\sqrt{N_n}} \sum_{t=b_n+1}^{n} \frac{u_{t-1} - u_t}{\eta_t} = o_p(1).$$

Proof. Write $a_t = \eta_t^{-1}$. Abel summation gives

$$\sum_{t=b_n+1}^{n} a_t(u_{t-1} - u_t) = a_{b_n+1}u_{b_n} - a_n u_n + \sum_{t=b_n+1}^{n-1} (a_{t+1} - a_t)u_t. \tag{26}$$

For late indices, (25) implies

$$a_t\mathbb{E}\,\|u_t\| = O(t^{\alpha - q/2}), \qquad (a_{t+1} - a_t)\mathbb{E}\,\|u_t\| = O(t^{\alpha - 1 - q/2}).$$

Choose a fixed index after which the moment bound holds and split every term in (26) at that index. The finitely many early terms, including a lower endpoint in this prefix, are bounded by a finite random variable and are $o_p(\sqrt{N_n})$. The expectation of the norm of the remaining terms is $O\{1 + n^{(\alpha - q/2)_+} \log n\}$, where the logarithm is needed only at the zero exponent. Since $q > 2\alpha - 1$, we have $\alpha - q/2 < 1/2$. Condition (2) gives $N_n \ge cn$ for all large n, so division by $\sqrt{N_n}$ yields a quantity tending to zero. Markov's inequality for the late terms and the pathwise prefix bound prove the claim. □

Lemma C.23. *Let $1 < p \le 2$ and let $D_t \in \mathbb{R}^d$ be martingale differences with $\mathbb{E}\|D_t\|^p < \infty$. There is a constant $C_{p,d}$ such that*

$$\mathbb{E}\left\|\sum_{t=k+1}^{k+N} D_t\right\|^p \le C_{p,d} \sum_{t=k+1}^{k+N} \mathbb{E}\|D_t\|^p \tag{27}$$

for all deterministic k, N.

Proof. For each coordinate, the martingale von Bahr–Esseen inequality gives

$$\mathbb{E}\left|\sum_t D_{t,j}\right|^p \le C_p \sum_t \mathbb{E}|D_{t,j}|^p;$$

see Hall & Heyde (1980, Chapter 2). Since $1 < p \le 2$,

$$\|x\|_2^p \le \sum_{j=1}^{d} |x_j|^p, \qquad \sum_{j=1}^{d} |x_j|^p \le d^{1-p/2} \|x\|_2^p.$$

Summing the coordinate inequalities and applying these norm relations proves (27). □

Proof of the exact reduction

Proof of proposition C.1. Because $\mathrm{D} = \mathrm{PB}(\mathrm{I}_r - \mathrm{B})^{-1}$ and B commutes with $(\mathrm{I}_r - \mathrm{B})^{-1}$,

$$\begin{aligned} \mathrm{P} + \mathrm{D} &= \mathrm{P}\{\mathrm{I}_r + \mathrm{B}(\mathrm{I}_r - \mathrm{B})^{-1}\} \\ &= \mathrm{P}(\mathrm{I}_r - \mathrm{B})^{-1}. \end{aligned} \tag{28}$$

Using $e_t = e_{t-1} - \eta_t \mathrm{P} v_t$ and (28),

$$\begin{aligned} u_t &= e_t - \eta_t \mathrm{D} v_t \\ &= e_{t-1} - \eta_t (\mathrm{P} + \mathrm{D}) v_t \\ &= e_{t-1} - \eta_t \mathrm{P}(\mathrm{I}_r - \mathrm{B})^{-1}\{\mathrm{B} v_{t-1} + \mathrm{C} G_t\} \\ &= e_{t-1} - \eta_t \mathrm{D} v_{t-1} - \eta_t \mathrm{K} G_t. \end{aligned}$$

Since $u_{t-1} = e_{t-1} - \eta_{t-1} \mathrm{D} v_{t-1}$,

$$e_{t-1} - \eta_t \mathrm{D} v_{t-1} = u_{t-1} + (\eta_{t-1} - \eta_t) \mathrm{D} v_{t-1}.$$

Substitution yields (17). □

Proof of asymptotic linearity

Proof of proposition C.2. By Lyapunov's inequality and (10),

$$\mathbb{E}\left[\|e_{t-1}\|^{1+\nu} \mathbf{1}(\theta_{t-1} \in \mathcal{U}) \right] \le \{\mathbb{E}\|e_{t-1}\|^2\}^{(1+\nu)/2} \le C t^{-\varrho(1+\nu)/2}.$$

Hence

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbb{E}\left[\|\mathfrak{r}(e_{t-1})\| \mathbf{1}(\theta_{t-1} \in \mathcal{U}) \right] \le \frac{C}{\sqrt{n}} \sum_{t=1}^{n} t^{-\varrho(1+\nu)/2} = o(1)$$

when $\varrho(1+\nu) > 1$. A finite initial segment is negligible. Markov's inequality now gives (11). □

Proof of Theorem 3.4. Let $e_t = \theta_t - \theta_\star$.

Step 1: transformed-state rate. By Assumption 3.3,

$$\begin{aligned} \mathbb{E}\|u_t\|^2 &\le 2\mathbb{E}\|e_t\|^2 + 2\eta_t^2 \|\mathrm{D}\|^2 \mathbb{E}\|v_t\|^2 \\ &= O(t^{-q_u}), \qquad q_u = \min(\varrho, 2\alpha). \end{aligned} \tag{29}$$

Because $\varrho > 2\alpha - 1$ and $2\alpha > 2\alpha - 1$, $q_u > 2\alpha - 1$. Thus lemma C.22 applies.

Step 2: exact decomposition. Use the mean remainder $\mathfrak{r}$ defined in Section 2 and let

$$\Delta\xi_t = \xi_t(\theta_{t-1}) - \xi_t(\theta_\star).$$

Then $G_t = \mathrm{A} e_{t-1} + \mathfrak{r}(e_{t-1}) + \xi_t^\star + \Delta\xi_t$ globally. Combining this identity with (17) gives

$$\begin{aligned} \mathrm{A} e_{t-1} = \mathrm{K}^{-1} \frac{u_{t-1} - u_t}{\eta_t} - \xi_t^\star - \Delta\xi_t - \mathfrak{r}(e_{t-1}) \\ + \frac{\eta_{t-1} - \eta_t}{\eta_t} \mathrm{K}^{-1} \mathrm{D} v_{t-1}. \end{aligned} \tag{30}$$

Step 3: localization. For fixed m, let $\tau_m = \inf\{t \geq m+1 : \theta_{t-1} \notin \mathcal{U}\}$, with $\inf \varnothing = \infty$. Then $\{t < \tau_m\} \in \mathcal{F}_{t-1}$ and

$$\mathbb{P}(\tau_m < \infty) \longrightarrow 0 \quad \text{as } m \to \infty. \tag{31}$$

Define stopped terms $\Delta\xi_t^{(m)} = \mathbf{1}\{t \geq m+1,\ t < \tau_m\}\Delta\xi_t$ and $\mathfrak{r}_{t-1}^{(m)} = \mathbf{1}\{t \geq m+1,\ t < \tau_m\}\mathfrak{r}(e_{t-1})$. Both are zero for $t \leq m$, so all localized moment bounds apply only after the finite prefix. For fixed m, finite terms before m are $o_p(\sqrt{N_n})$. Bounds for the stopped sums transfer to the original sums by first letting $n \to \infty$ and then $m \to \infty$ in (31).

Step 4: Abel and filter-variation terms. By (29) and lemma C.22,

$$\frac{1}{\sqrt{N_n}} \sum_{t=b_n+1}^{n} \mathrm{K}^{-1} \frac{u_{t-1} - u_t}{\eta_t} = o_p(1).$$

By lemma C.21 and the filter-state bound in Assumption 3.3,

$$\begin{aligned}\mathbb{E}\left\| \frac{1}{\sqrt{N_n}} \sum_{t=b_n+1}^{n} \frac{\eta_{t-1} - \eta_t}{\eta_t} \mathrm{K}^{-1} \mathrm{D} v_{t-1} \right\| &\leq \frac{C}{\sqrt{N_n}} \sum_{t=2}^{n} \frac{1}{t} \\ &= O\left(\frac{\log n}{\sqrt{n}}\right) = o(1).\end{aligned}$$

Step 5: local score variation. For fixed m, the stopped sequence $\Delta\xi_t^{(m)}$ is a martingale difference. Orthogonality gives

$$\mathbb{E}\left\| \frac{1}{\sqrt{N_n}} \sum_{t=\max(b_n+1,m+1)}^{n} \Delta\xi_t^{(m)} \right\|^2 \leq \frac{1}{N_n} \sum_{t=b_n+1}^{n} \mathbb{E}\left[\omega_\xi(\|e_{t-1}\|)\mathbf{1}(t < \tau_m)\right].$$

Because $\varrho > 0$, $e_t \to 0$ in probability. The function ω_ξ is bounded and tends to zero at the origin, so the expectation in the last display tends to zero. Cesàro averaging over an interval satisfying (2) therefore gives an $o(1)$ bound. The stopped score-variation sum is $o_p(1)$, and localization transfers the result to the original sum.

Step 6: nonlinear mean remainder. For fixed m, $N_n \geq cn$ and (11) give

$$\left\| \frac{1}{\sqrt{N_n}} \sum_{t=\max(b_n+1,m+1)}^{n} \mathfrak{r}_{t-1}^{(m)} \right\| \leq \frac{C}{\sqrt{n}} \sum_{t=1}^{n} \|\mathfrak{r}(e_{t-1})\|\, \mathbf{1}(\theta_{t-1} \in \mathcal{U}) \xrightarrow{p} 0.$$

Localization gives the same conclusion for the original remainder sum.

Step 7: summation. Summing (30) over $t = b_n+1, \ldots, n$ and using Steps 3–6 yields

$$\frac{1}{\sqrt{N_n}} \sum_{t=b_n+1}^{n} e_{t-1} = -\mathrm{A}^{-1} \frac{1}{\sqrt{N_n}} \sum_{t=b_n+1}^{n} \xi_t^\star + o_p(1). \tag{32}$$

The difference between the sums of e_t and e_{t-1} is $e_n - e_{b_n}$. If $b_n \to \infty$, the mean-square rate makes both endpoints $o_p(\sqrt{N_n})$; if b_n is bounded along a subsequence, $e_{b_n} = O_p(1)$ and the same conclusion holds. Replacing e_{t-1} by e_t in (32) proves (12).

Step 8: normal limit. The sum $\sum_{t=b_n+1}^{n} \xi_t^\star$ has the distribution of a sum of N_n i.i.d. centered random vectors with covariance S. Since $N_n \to \infty$ and the second moment is finite, the multivariate central limit theorem, (12), and Slutsky's theorem yield the stated normal limit. □

Proof of corollary C.3. Subtracting the assumed same-window batch expansion from (12) proves corollary C.3. For the full-sample statement, multiply (12) by $\sqrt{n/N_n}$ to obtain

$$\sqrt{n}(\bar{\theta}_{n,b} - \theta_\star) = -\mathrm{A}^{-1} \frac{\sqrt{n}}{N_n} \sum_{t=b_n+1}^{n} \xi_t^\star + o_p(1).$$

Subtract the full-sample leading term. The difference equals

$$\frac{\sqrt{n}}{N_n}\sum_{t=b_n+1}^{n}\xi_t^\star-\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\xi_t^\star$$
$$=\left(\frac{n}{N_n}-1\right)\frac{1}{\sqrt{n}}\sum_{t=b_n+1}^{n}\xi_t^\star-\frac{1}{\sqrt{n}}\sum_{t=1}^{b_n}\xi_t^\star.$$

If $b_n = o(n)$, the first term is $O_p(b_n/n) = o_p(1)$ and the second is $O_p\{(b_n/n)^{1/2}\} = o_p(1)$. Comparing with the assumed full-sample batch expansion proves corollary C.3. For the filter-equivalence assertion, apply Theorem 3.4 to each recursion. Both expansions have the same leading term $-\mathrm{A}^{-1}N_n^{-1/2}\sum_{t=b_n+1}^{n}\xi_t^\star$. Subtracting them gives corollary C.3. □

C.7.2 Proofs for the functional limits and self-normalization

Integrated remainder bounds

Lemma C.24. *Let u_t be finite almost surely for every t, and suppose that $\mathbb{E}\|u_t\|^2 \le Ct^{-q}$ for all sufficiently large t, where $q > 2\alpha - 1$ and $\eta_t = \eta_0 t^{-\alpha}$, $0 < \alpha < 1$. For $H_0 = 0$ and $H_k = \sum_{t=1}^{k}\eta_t^{-1}(u_{t-1} - u_t)$,*

$$\frac{1}{n^2}\sum_{k=0}^{n-1}\|H_k\|^2 \xrightarrow{p} 0.$$

If, in addition, all finite-prefix second moments are finite, the same convergence holds in expectation. For retained intervals with $N_n \asymp n$, the corresponding Abel sum divided by $\sqrt{N_n}$ converges to zero in probability under the same finite-prefix convention.

Proof. Set $a_t = \eta_t^{-1}$ and choose a deterministic m after which the second-moment bound holds. For $k > m$, Abel summation writes H_k as a finite-prefix random constant plus $-a_k u_k + \sum_{t=m}^{k-1}(a_{t+1} - a_t)u_t$. The L^2 norm of this latter expression is bounded by a constant times 1, $\log(k+1)$, or k^δ, according as $\delta = \alpha - q/2$ is negative, zero, or positive. Since $\delta < 1/2$, its squared L^2 norms sum to $o(n^2)$. The squared finite-prefix contribution divided by n converges to zero almost surely, even when that random constant has infinite expectation. The finitely many terms with $k \le m$ contribute $o(1)$ almost surely. The inequality $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$ and Markov's inequality for the late part prove convergence in probability. If all prefix second moments are finite, the argument can also be integrated. For a moving interval, use

$$\sum_{t=b_n+1}^{n}a_t(u_{t-1}-u_t) = a_{b_n+1}u_{b_n} - a_n u_n + \sum_{t=b_n+1}^{n-1}(a_{t+1}-a_t)u_t.$$

Split early indices pathwise and bound the remaining terms in L^1 by $O\{1 + n^{\delta_+}\log(n+1)\} = o(\sqrt{n})$. □

Lemma C.25. *Let D_t be square-integrable vector martingale differences and put $a_t = \mathbb{E}\,\|D_t\|^2$. If $a_t \to 0$, then*

$$\mathbb{E}\int_0^1\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}D_t\right\|^2 ds \longrightarrow 0.$$

Proof. Orthogonality of martingale differences gives

$$\mathbb{E}\left\|\sum_{t=1}^{k}D_t\right\|^2 = \sum_{t=1}^{k}a_t.$$

Therefore

$$\mathbb{E}\int_0^1\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}D_t\right\|^2 ds = \frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{t=1}^{k}a_t$$

$$\leq \frac{1}{n}\sum_{t=1}^{n} a_t \longrightarrow 0$$

by Cesàro convergence. □

Proof of the functional central limit theorem

Proof of Theorem 3.7. Let $e_t = \theta_t - \theta_\star$ and $q_u = \min(\varrho, 2\alpha)$. As in (29),

$$\mathbb{E}\left\|u_t\right\|^2 = O(t^{-q_u}), \qquad q_u > 2\alpha - 1. \tag{33}$$

Define the target-score process

$$\mathbb{Z}_n(s) = \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor} \xi_t^\star.$$

The multivariate invariance principle gives

$$\{\mathbb{Z}_n, \mathbb{Z}_n(1)\} \Longrightarrow \{\mathrm{S}^{1/2}W_d, \mathrm{S}^{1/2}W_d(1)\} \tag{34}$$

in $D([0,1];\mathbb{R}^d)\times\mathbb{R}^d$ (Billingsley, 1999). Since the limit paths are continuous, the embedding into $\mathbb{H}\times\mathbb{R}^d$ is continuous at the limit.

We next show that all terms other than $-\mathrm{A}^{-1}\mathbb{Z}_n$ vanish in $\mathbb{H}$. For fixed m, use the stopping time τ_m and the stopped score and remainder terms defined in the proof of Theorem 3.4. Finite terms before m have $\mathbb{H}$ norm $O_p(n^{-1/2})$. On $\{\tau_m = \infty\}$ the stopped and original terms agree after time m, and (31) permits $m \to \infty$ after $n \to \infty$.

For the Abel term, lemma C.24 and (33) yield

$$\int_0^1 \left\| \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor} \mathrm{K}^{-1}\frac{u_{t-1}-u_t}{\eta_t}\right\|^2 ds \xrightarrow{p} 0.$$

For the stepsize-variation term, lemma C.21 and the state bound give, uniformly in k,

$$\left\{\mathbb{E}\left\|\sum_{t=2}^{k}\frac{\eta_{t-1}-\eta_t}{\eta_t}\mathrm{K}^{-1}\mathrm{D}v_{t-1}\right\|^2\right\}^{1/2} \leq C\sum_{t=2}^{k} t^{-1} \leq C\log(k+1).$$

Consequently its expected squared $\mathbb{H}$ norm is bounded by $Cn^{-2}\sum_{k\leq n}\log^2(k+1) = o(1)$.

For the stopped score variation, set $D_t = \Delta\xi_t^{(m)}$. It is a martingale difference and

$$a_t = \mathbb{E}\left\|D_t\right\|^2 \leq \mathbb{E}\{\omega_\xi(\|e_{t-1}\|)\} \longrightarrow 0,$$

because $e_t \to 0$ in probability and ω_ξ is bounded and continuous at zero. Lemma C.25 therefore applies. For the stopped mean remainder,

$$\sup_{0\leq s\leq 1}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}\mathfrak{r}_{t-1}^{(m)}\right\| \leq \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\|\mathfrak{r}(e_{t-1})\|\,\mathbf{1}(\theta_{t-1}\in\mathcal{U}) \xrightarrow{p} 0$$

by (11). Summing (30) up to $\lfloor ns\rfloor$ now gives

$$\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor n\cdot\rfloor} e_{t-1} + \mathrm{A}^{-1}\mathbb{Z}_n\right\|_{\mathbb{H}} \xrightarrow{p} 0. \tag{35}$$

It remains to replace e_{t-1} by e_t. The difference at time s is $n^{-1/2}\{e_{\lfloor ns\rfloor} - e_0\}$. Since $e_t = e_0 - \sum_{j=1}^{t}\eta_j\mathrm{P}v_j$ gives finite second moments at every fixed time, the late mean-square rate implies $\sup_t \mathbb{E}\left\|e_t\right\|^2 < \infty$,

$$\mathbb{E}\int_0^1 \frac{1}{n}\left\|e_{\lfloor ns\rfloor} - e_0\right\|^2 ds \leq \frac{C}{n^2}\sum_{k=0}^{n-1}\{\mathbb{E}\left\|e_k\right\|^2 + \mathbb{E}\left\|e_0\right\|^2\} = O(n^{-1}).$$

Thus (35) holds with $\mathbb{U}_n$. The endpoint relation is Theorem 3.4 with $b_n = 0$. Combining these two approximations with (34) proves the joint convergence in Theorem 3.7. □

Proof of proposition C.8. We strengthen each integrated remainder bound to a uniform one. Abel summation gives

$$\max_{k\le n}\|H_k\| \le a_1\|u_0\| + \max_{k\le n} a_k\|u_k\| + \sum_{t=1}^{n-1}(a_{t+1}-a_t)\|u_t\|.$$

The second term is $o_p(\sqrt{n})$ by (20). The expectation of the last term is

$$O\left(\sum_{t=1}^{n-1} t^{\alpha-1-q_u/2}\right) = o(\sqrt{n})$$

because $q_u > 2\alpha - 1$. Hence the Abel remainder is uniformly $o_p(\sqrt{n})$. The stepsize-variation remainder is uniformly $o_p(\sqrt{n})$ by the same $O(\log n)$ expectation bound.

For the stopped score variation, Doob's inequality gives

$$\mathbb{E}\max_{k\le n}\left\|\sum_{t=1}^{k}\Delta\xi_t^{(m)}\right\|^2 \le 4\sum_{t=1}^{n}\mathbb{E}\left\|\Delta\xi_t^{(m)}\right\|^2 = o(n).$$

The accumulated remainder condition already controls the supremum of all its partial sums. Finally,

$$\mathbb{P}\left(\max_{k\le n}\|e_k\| > \varepsilon\sqrt{n}\right) \le \frac{1}{\varepsilon^2 n}\sum_{k=1}^{n}\mathbb{E}\|e_k\|^2 \longrightarrow 0,$$

so replacing e_{t-1} by e_t is uniformly negligible. Localization is handled as in the proof of Theorem 3.7. The score process satisfies Donsker's theorem, and all other terms vanish uniformly, proving proposition C.8.

For the stated sufficient condition, a union bound and Markov's inequality give

$$\begin{aligned}\mathbb{P}\left(\max_{k\le n}\eta_k^{-1}\|u_k\| > \varepsilon\sqrt{n}\right) &\le C\varepsilon^{-p}n^{-p/2}\sum_{k=1}^{n}k^{p\alpha}\mathbb{E}\|u_k\|^p \\ &\le C\varepsilon^{-p}n^{-p/2}\sum_{k=1}^{n}k^{p(\alpha-q_u/2)}.\end{aligned}$$

If the exponent in the final sum exceeds -1, the bound is $O\{n^{1-p(q_u+1-2\alpha)/2}\} = o(1)$ by proposition C.8; if it equals or is below -1, the conclusion is immediate. This proves (20). □

Proof of self-normalized inference

Proof of Theorem 3.8. Let $h_n(s) = \lfloor ns\rfloor/n$. By the definition of $\mathbb{U}_n$ and (8),

$$\widehat{\mathrm{V}}_n = \int_0^1 \{\mathbb{U}_n(s) - h_n(s)\mathbb{U}_n(1)\}\{\mathbb{U}_n(s) - h_n(s)\mathbb{U}_n(1)\}^{\mathsf{T}} ds. \tag{36}$$

Indeed, on $[k/n, (k+1)/n)$ the vector inside braces equals $n^{-1/2}\{\sum_{t=1}^{k}\theta_t - (k/n)\sum_{t=1}^{n}\theta_t\}$, and the target $\theta_\star$ cancels. Summing the integrals over the grid gives (8).

The map

$$(x,z) \longmapsto \int_0^1 \{x(s) - sz\}\{x(s) - sz\}^{\mathsf{T}} ds$$

is continuous from $\mathbb{H}\times\mathbb{R}^d$ to the space of symmetric matrices. Moreover, $\sup_s |h_n(s) - s| \le n^{-1}$, while Theorem 3.7 implies $\mathbb{U}_n(1) = O_p(1)$. Replacing h_n by s in (36) therefore changes the matrix by $o_p(1)$. The continuous mapping theorem proves the joint convergence asserted in Theorem 3.8.

For a fixed contrast a, the process $a^{\mathsf{T}}\Lambda W_d$ is a one-dimensional Brownian motion multiplied by the positive constant $(a^{\mathsf{T}}\Sigma a)^{1/2}$. The same constant multiplies the numerator and denominator in (13), and hence cancels. The Brownian endpoint is independent of the bridge; conditional on the strictly

positive bridge norm, the limiting ratio has a nondegenerate normal distribution. Its distribution is therefore continuous, which justifies quantile inversion and the coverage of (9).

For the vector result, $\mathrm{R}_0 \Lambda W_d$ has the same law as $\Omega^{1/2} W_q$, where $\Omega = \mathrm{R}_0 \Sigma \mathrm{R}_0^{\mathsf{T}}$. Its self-normalizer has limit $\Omega^{1/2} \mathrm{J}_q \Omega^{1/2}$. The two factors of $\Omega^{1/2}$ cancel from the quadratic form, yielding (14).

It remains to verify that J_q is positive definite almost surely. If it were singular, continuity of the Brownian bridge would imply that its values lie in a proper linear subspace for every $s \in [0, 1]$. Choose distinct points $0 < s_1 < \cdots < s_q < 1$. The $q \times q$ matrix whose columns are $W_q(s_j) - s_j W_q(1)$ has a nondegenerate Gaussian density because the Brownian-bridge covariance matrix at these points is positive definite. Its determinant is therefore nonzero almost surely, a contradiction. Thus J_q is positive definite almost surely. Since Ω is positive definite, the smallest eigenvalue of $\mathrm{R}_0 \widehat{\mathrm{V}}_n \mathrm{R}_0^{\mathsf{T}}$ converges in distribution to a strictly positive random variable. The sample matrix is positive semidefinite, so its probability of singularity tends to zero. The scalar normalizer is positive with probability tending to one by the same argument. Extending a pivot or withholding its report on these vanishing-probability events changes neither its weak limit nor its limiting coverage. □

C.7.3 Transient stabilization and primitive mean-square stability

Finite stabilization corrections

Proposition C.26. *Suppose the distributional, trajectory, and burn-in conditions of Theorem 3.4 hold for an adapted trajectory with updates*

$$v_t = \mathrm{B} v_{t-1} + \mathrm{C} G(\theta_{t-1}, Z_t) + \zeta_t, \qquad \theta_t = \theta_{t-1} - \eta_t \mathrm{P} v_t + q_t.$$

Assume all iterates and corrections are finite almost surely, and $q_t = \zeta_t = 0$ for $t > T$, where $T < \infty$ almost surely. The terminal conclusion of Theorem 3.4 and the joint integrated limit of Theorem 3.7 remain valid. Consequently, Theorem 3.8 also applies. If the additional maximal condition of proposition C.8 holds, its path-space conclusion is unchanged as well.

Proof. The transformed recursion acquires the term $q_t - \eta_t \mathrm{P}(\mathrm{I}_r - \mathrm{B})^{-1} \zeta_t$. In the estimating-equation decomposition this becomes

$$c_t = \eta_t^{-1} \mathrm{K}^{-1} q_t - \mathrm{K}^{-1} \mathrm{P}(\mathrm{I}_r - \mathrm{B})^{-1} \zeta_t.$$

For the full partial sums,

$$\frac{1}{\sqrt{n}} \max_{k \le n} \left\| \sum_{t=1}^{k} c_t \right\| \le \frac{1}{\sqrt{n}} \sum_{t=1}^{T} \|c_t\| \longrightarrow 0 \quad \text{almost surely.}$$

The analogous retained-sample bound has denominator $\sqrt{N_n}$ and is also negligible. This uniform bound implies negligibility in $L^2([0, 1])$ and at the endpoint. Apply lemmas C.22 and C.24 and predictable localization to the remaining terms. For the change from e_{t-1} to e_t, split at a deterministic index after which the mean-square rate holds. The early values are finite almost surely, and the late values satisfy $n^{-2} \sum_{k \le n} \mathbb{E}\|e_k\|^2 = o(1)$ after that split. Thus the joint integrated and endpoint approximation remains valid. No moment bound on the finite corrections themselves is needed. The last assertion uses the same uniform approximation and the separately assumed maximal condition. The proposition does not prove that a particular reset or truncation scheme stabilizes, nor does it cover persistently active projection. □

This proposition covers expanding truncation or reset mechanisms after it has been shown that only finitely many parameter and filter-state corrections occur. It does not cover a projection that remains active with nonnegligible frequency; persistent projection requires a separate constrained-inference analysis.

A deterministic comparison lemma

Lemma C.27. *Let $a_t = a_0 t^{-\alpha}$ with $0 < \alpha < 1$. If a nonnegative sequence satisfies*

$$x_t \le (1 - c a_t) x_{t-1} + C a_t^2 \tag{37}$$

for all sufficiently large t and constants $c, C > 0$, then $x_t = O(a_t)$.

Proof. Choose $M > 2C/c$. Since

$$\frac{a_{t-1}}{a_t} = \left(\frac{t}{t-1}\right)^{\alpha} = 1 + O(t^{-1}) = 1 + o(a_t),$$

there is t_0 such that $Ma_{t-1}(1 - ca_t) + Ca_t^2 \le Ma_t$ for all $t \ge t_0$. Increase M if necessary so that $x_{t_0} \le Ma_{t_0}$. Induction using (37) gives $x_t \le Ma_t$ for every $t \ge t_0$. □

Proof of the primitive rate theorem

Proof of Theorem C.5. Let $e_t = \theta_t - \theta_\star$. For the scalar filter in (18), define

$$\mathrm{D} = \frac{\beta}{1-\beta}\mathrm{P}, \qquad \mathrm{K} = \frac{c}{1-\beta}\mathrm{P}, \qquad u_t = e_t - \eta_t \mathrm{D}m_t. \tag{38}$$

By proposition C.1,

$$u_t = u_{t-1} - \eta_t \mathrm{K}G(\theta_{t-1}, Z_t) + (\eta_{t-1} - \eta_t)\mathrm{D}m_{t-1}. \tag{39}$$

Write

$$U_t = \mathbb{E}\left\|u_t\right\|^2, \qquad M_t = \mathbb{E}\left\|m_t\right\|^2.$$

All constants below may change from line to line but do not depend on t.

Step 1: convert the primitive bounds to unfiltered-score bounds. Because K is nonsingular, the growth and noise bounds in Assumption C.4 imply

$$\|g(\theta_\star + e)\| \le C\|e\|, \tag{40}$$

$$\mathbb{E}\|\xi(\theta_\star + e, Z_1)\|^2 \le C(1 + \|e\|^2). \tag{41}$$

Step 2: momentum-state recursion. Conditionally on $\mathcal{F}_{t-1}$, $m_t = \beta m_{t-1} + cg(\theta_{t-1}) + c\xi_t(\theta_{t-1})$. Choose $\varepsilon > 0$ so that $q_0 = (1+\varepsilon)\beta^2 < 1$. Young's inequality, (40), and (41) yield

$$\mathbb{E}(\|m_t\|^2 \mid \mathcal{F}_{t-1}) \le q_0\|m_{t-1}\|^2 + C\{1 + \|e_{t-1}\|^2\}.$$

From (38), $e_{t-1} = u_{t-1} + \eta_{t-1}\mathrm{D}m_{t-1}$, and hence

$$\|e_{t-1}\|^2 \le 2\|u_{t-1}\|^2 + 2\eta_{t-1}^2\|\mathrm{D}\|^2\|m_{t-1}\|^2. \tag{42}$$

Since $\eta_t \to 0$, the last contribution can be absorbed into some $q \in (q_0, 1)$ after a finite deterministic index. Taking expectations gives

$$M_t \le qM_{t-1} + CU_{t-1} + C. \tag{43}$$

Step 3: lower bound the transformed drift. Suppress the index $t-1$ and write $e = e_{t-1}$, $u = u_{t-1}$, and $m = m_{t-1}$. Since $u = e - \eta_{t-1}\mathrm{D}m$,

$$\begin{aligned}\langle u, \mathrm{K}g(\theta_{t-1})\rangle &= \langle e, \mathrm{K}g(\theta_{t-1})\rangle - \eta_{t-1}\langle \mathrm{D}m, \mathrm{K}g(\theta_{t-1})\rangle \\ &\ge \mu\|e\|^2 - C\eta_{t-1}\|m\|\|e\| \\ &\ge \frac{\mu}{2}\|e\|^2 - C\eta_{t-1}^2\|m\|^2.\end{aligned}$$

Also, from $u = e - \eta_{t-1}\mathrm{D}m$,

$$\|e\|^2 \ge \frac{1}{2}\|u\|^2 - \eta_{t-1}^2\|\mathrm{D}\|^2\|m\|^2.$$

Therefore

$$\langle u, \mathrm{K}g(\theta_{t-1})\rangle \ge c_0\|u\|^2 - C\eta_{t-1}^2\|m\|^2 \tag{44}$$

for $c_0 = \mu/4$.

Step 4: conditional transformed-state recursion. Set

$$d_t = (\eta_{t-1} - \eta_t)\mathrm{D}m_{t-1}, \qquad h_{t-1} = \mathrm{K}g(\theta_{t-1}).$$

Using (39) and the martingale-difference property,

$$\mathbb{E}(\|u_t\|^2 \mid \mathcal{F}_{t-1}) = \|u_{t-1} - \eta_t h_{t-1} + d_t\|^2 + \eta_t^2 \mathbb{E}\{\|\mathrm{K}\xi_t(\theta_{t-1})\|^2 \mid \mathcal{F}_{t-1}\}. \tag{45}$$

We bound the terms in order. By (44) and Assumption C.4,

$$\begin{aligned}\|u_{t-1} - \eta_t h_{t-1}\|^2 &= \|u_{t-1}\|^2 - 2\eta_t \langle u_{t-1}, h_{t-1}\rangle + \eta_t^2 \|h_{t-1}\|^2 \\ &\le (1 - 2c_0\eta_t)\|u_{t-1}\|^2 + C\eta_t\eta_{t-1}^2 \|m_{t-1}\|^2 + C\eta_t^2 \|e_{t-1}\|^2 .\end{aligned}$$

Using (42) and $\eta_{t-1}/\eta_t \to 1$, the last line is, for all large t, bounded by

$$(1 - c_1\eta_t)\|u_{t-1}\|^2 + C\eta_t^3 \|m_{t-1}\|^2 \tag{46}$$

for some $c_1 > 0$.

Next, lemma C.21 gives

$$\|d_t\|^2 \le C\frac{\eta_t^2}{t^2}\|m_{t-1}\|^2 .$$

For $x_t = u_{t-1} - \eta_t h_{t-1}$, Young's inequality with coefficient $c_1\eta_t/2$ gives

$$\begin{aligned}2\langle x_t, d_t\rangle &\le \frac{c_1\eta_t}{2}\|x_t\|^2 + \frac{2}{c_1\eta_t}\|d_t\|^2 \\ &\le \frac{c_1\eta_t}{2}\|x_t\|^2 + C\frac{\eta_t}{t^2}\|m_{t-1}\|^2 .\end{aligned}$$

Because $\alpha < 1$, $\eta_t/t^2 = O(\eta_t^3)$. The small multiple of $\|x_t\|^2$ can be absorbed into the negative drift in (46). Finally, Assumption C.4 and (42) imply

$$\eta_t^2 \mathbb{E}\{\|\mathrm{K}\xi_t(\theta_{t-1})\|^2 \mid \mathcal{F}_{t-1}\} \le C\eta_t^2 + C\eta_t^2\|u_{t-1}\|^2 + C\eta_t^2\eta_{t-1}^2\|m_{t-1}\|^2 .$$

The $\eta_t^2\|u\|^2$ term is again absorbed into the negative $\eta_t\|u\|^2$ drift. Substitution into (45) and expectation yield

$$U_t \le (1 - c\eta_t)U_{t-1} + C\eta_t^2 + C\eta_t^3 M_{t-1} \tag{47}$$

for all sufficiently large t.

Step 5: uniform boundedness of the coupled system. Fix t_0 after which (43) and (47) hold. Iterating (43) gives

$$M_t \le C_0 + C\max_{t_0 \le s < t} U_s, \qquad t > t_0, \tag{48}$$

where C_0 depends on the state at t_0 but not on t.

For $T \ge t_0$, set $\overline{U}_T = \max_{t_0 \le s \le T} U_s$. If the maximum is attained at t_0, then $\overline{U}_T = U_{t_0}$. Otherwise choose $t_\star \in \{t_0 + 1, \dots, T\}$ with $U_{t_\star} = \overline{U}_T$. Equations (47) and (48) imply

$$\overline{U}_T \le (1 - c\eta_{t_\star})\overline{U}_T + C\eta_{t_\star}^2 + C\eta_{t_\star}^3(C_0 + C\overline{U}_T).$$

Increase t_0 so that the last coefficient of $\overline{U}_T$ is at most $c\eta_{t_\star}/2$. Then

$$\overline{U}_T \le C\{\eta_{t_\star} + \eta_{t_\star}^2\} \le C\eta_{t_0}.$$

Thus $\sup_t U_t < \infty$, without using square summability of the stepsizes. Equation (48) then gives $\sup_t M_t < \infty$.

Step 6: sharp $O(\eta_t)$ rate. With M_t uniformly bounded, (47) reduces to $U_t \le (1 - c\eta_t)U_{t-1} + C\eta_t^2$. By lemma C.27, $U_t = O(\eta_t)$. Finally, from $e_t = u_t + \eta_t \mathrm{D}m_t$,

$$\mathbb{E}\|e_t\|^2 \le 2U_t + 2\eta_t^2\|\mathrm{D}\|^2 M_t = O(\eta_t).$$

Together with $\sup_t M_t < \infty$, this proves the moment conclusions.

Step 7: almost-sure convergence when $\alpha > 1/2$. The conditional calculation in Step 4 also gives

$$\mathbb{E}(\|u_t\|^2 \mid \mathcal{F}_{t-1}) \le (1 - c\eta_t)\|u_{t-1}\|^2 + C\eta_t^2 + C\eta_t^3\|m_{t-1}\|^2.$$

The additive terms have summable expectations. The nonnegative almost-supermartingale argument therefore gives convergence of $\|u_t\|^2$ and $\sum_t \eta_t\|u_{t-1}\|^2 < \infty$ almost surely. Since $\sum_t \eta_t = \infty$, the limit is zero. Also $\mathbb{E}\sum_t \eta_t^2\|m_t\|^2 < \infty$, so $\eta_t m_t \to 0$ almost surely. Hence $e_t = u_t + \eta_t \mathrm{D} m_t \to 0$ almost surely, as claimed. □

Proof of Theorem C.6. Let $e_t = \theta_t - \theta_\star$ and define the pre-projection point and projection correction by

$$x_t = \theta_{t-1} - \eta_t \mathrm{P} m_t, \qquad q_t = \Pi_{\mathcal{C}}(x_t) - x_t.$$

Thus $\theta_t = x_t + q_t$. Set

$$\mathrm{D} = \frac{\beta}{1-\beta}\mathrm{P}, \qquad u_t = e_t - \eta_t \mathrm{D} m_t, \qquad d_t = (\eta_{t-1} - \eta_t)\mathrm{D} m_{t-1}.$$

Step 1: a uniform second-moment bound for the momentum state. Projection gives $\theta_t \in \mathcal{C}$ for every t. Compactness of $\mathcal{C}$, the drift and noise bounds in Theorem C.6, and nonsingularity of K imply

$$\sup_{\theta\in\mathcal{C}} \mathbb{E}\,\|G(\theta, Z_1)\|^2 \le \left\|\mathrm{K}^{-1}\right\|^2 \sup_{\theta\in\mathcal{C}} \left\{\|\mathrm{K} g(\theta)\|^2 + \mathbb{E}\,\|\mathrm{K}\xi(\theta, Z_1)\|^2\right\} < \infty.$$

Conditionally on $\mathcal{F}_{t-1}$, Young's inequality with a constant $q_m \in (\beta^2, 1)$ therefore gives

$$\mathbb{E}(\|m_t\|^2 \mid \mathcal{F}_{t-1}) \le q_m \|m_{t-1}\|^2 + C. \tag{49}$$

Iteration proves $\sup_t \mathbb{E}\,\|m_t\|^2 < \infty$.

Step 2: projection preserves the transformed drift inequality. The scalar-filter identity used in (39), now including q_t, is

$$u_t = u_{t-1} - \eta_t \mathrm{K} G(\theta_{t-1}, Z_t) + d_t + q_t.$$

Euclidean projection and $\theta_\star \in \mathcal{C}$ imply the variational inequality

$$\langle q_t, e_t \rangle \le 0.$$

Let $a_t = u_{t-1} - \eta_t \mathrm{K} G(\theta_{t-1}, Z_t) + d_t$, so that $u_t = a_t + q_t$. Since $u_t = e_t - \eta_t \mathrm{D} m_t$,

$$\begin{aligned}
\|u_t\|^2 &= \|a_t\|^2 + 2\langle u_t, q_t\rangle - \|q_t\|^2 \\
&\le \|a_t\|^2 - 2\eta_t \langle \mathrm{D} m_t, q_t\rangle - \|q_t\|^2 \\
&\le \|a_t\|^2 + \eta_t^2 \|\mathrm{D} m_t\|^2 .
\end{aligned} \tag{50}$$

Thus projection contributes only a second-order state term.

Step 3: the mean-square rate is $O(\eta_t)$. Because $\theta_{t-1} \in \mathcal{C}$, the drift calculation in (44) and the stepsize calculation following (46) apply under the local drift and noise bounds in Theorem C.6. Before adding the projection correction they give, for all sufficiently large t,

$$\mathbb{E}(\|a_t\|^2 \mid \mathcal{F}_{t-1}) \le (1 - c_0\eta_t)\|u_{t-1}\|^2 + C\eta_t^2 + C\eta_t^3\|m_{t-1}\|^2 \tag{51}$$

for some $c_0 > 0$. Combining (50), (49), and (51) gives

$$\mathbb{E}(\|u_t\|^2 \mid \mathcal{F}_{t-1}) \le \mathbb{E}(\|a_t\|^2 \mid \mathcal{F}_{t-1}) + \eta_t^2 \|\mathrm{D}\|^2 \{q_m \|m_{t-1}\|^2 + C\}.$$

After reducing c_0 if necessary, this yields

$$\mathbb{E}(\|u_t\|^2 \mid \mathcal{F}_{t-1}) \le (1 - c\eta_t)\|u_{t-1}\|^2 + C\eta_t^2\{1 + \|m_{t-1}\|^2\}. \tag{52}$$

Taking expectations and using Step 1 gives

$$U_t \le (1 - c\eta_t)U_{t-1} + C\eta_t^2, \qquad U_t = \mathbb{E}\,\|u_t\|^2 .$$

By lemma C.27, $U_t = O(\eta_t)$. The identity $e_t = u_t + \eta_t \mathrm{D} m_t$ and Step 1 then give

$$\mathbb{E}\,\|e_t\|^2 \le 2U_t + 2\eta_t^2 \|\mathrm{D}\|^2 \mathbb{E}\,\|m_t\|^2 = O(\eta_t). \tag{53}$$

Step 4: the projection is eventually inactive. The condition $\alpha(1+\nu) > 1$ and $\nu \le 1$ imply $\alpha > 1/2$, so $\sum_t \eta_t^2 < \infty$. Step 1 gives

$$\sum_t \eta_t^2 \|m_{t-1}\|^2 < \infty \quad \text{almost surely.} \tag{54}$$

Apply the almost-supermartingale argument to (52): with $b_t = C\eta_t^2\{1 + \|m_{t-1}\|^2\}$, we have $\mathbb{E}\sum_t b_t < \infty$. More explicitly,

$$Y_t = \|u_t\|^2 + \mathbb{E}\left(\sum_{s=t+1}^{\infty} b_s \middle| \mathcal{F}_t \right)$$

is, after a finite index, a nonnegative supermartingale with conditional drift at most $-c\eta_t \|u_{t-1}\|^2$. Hence Y_t converges and $\sum_t \eta_t \|u_{t-1}\|^2 < \infty$ almost surely. Since $B_\infty = \sum_{s=1}^{\infty} b_s$ is integrable and measurable with respect to $\mathcal{F}_\infty = \sigma(\bigcup_t \mathcal{F}_t)$, martingale convergence gives $\mathbb{E}(B_\infty \mid \mathcal{F}_t) \to B_\infty$ almost surely. Therefore

$$\mathbb{E}\left(\sum_{s=t+1}^{\infty} b_s \middle| \mathcal{F}_t \right) = \mathbb{E}(B_\infty \mid \mathcal{F}_t) - \sum_{s=1}^{t} b_s \longrightarrow 0 \quad \text{almost surely,}$$

so $\|u_t\|^2$ converges. Since $\sum_t \eta_t = \infty$, its limit is zero. Equation (54) also gives $\eta_{t+1}\|m_t\| \to 0$ almost surely. Since $\eta_t/\eta_{t+1} \to 1$, also $\eta_t \|m_t\| \to 0$. Therefore $e_t = u_t + \eta_t \mathrm{D} m_t \to 0$ almost surely.

Let $\delta > 0$ satisfy $B(\theta_\star, \delta) \subset \mathcal{C}$. Almost surely, both $\|e_{t-1}\| < \delta/2$ and $\eta_t \|\mathrm{P} m_t\| < \delta/2$ hold for all sufficiently large t. Then $x_t \in B(\theta_\star, \delta)$, so $q_t = 0$. The projection is therefore active only finitely often.

Step 5: accumulated bias and inference. By the remainder bound in Theorem C.6, Lyapunov's inequality, and (53),

$$\mathbb{E}\,\|\mathfrak{r}(e_t)\| \le L_r\{\mathbb{E}\,\|e_t\|^2\}^{(1+\nu)/2} \le Ct^{-\alpha(1+\nu)/2}.$$

Consequently,

$$\mathbb{E}\left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \|\mathfrak{r}(e_{t-1})\| \right\} \longrightarrow 0$$

because $\alpha(1+\nu)/2 > 1/2$. Markov's inequality proves (11). Projection already confines the path to $\mathcal{C} \subset \mathcal{U}$; Steps 1 and 3 give the remaining parts of Assumption 3.3 with $\varrho = \alpha$. Finally, the projection corrections q_t vanish after a finite almost-sure time, so proposition C.26 transfers the asymptotic linear representation and normal limit in Theorem 3.4 to the projected recursion. □

C.7.4 Proofs for online covariance estimation

Lemma C.28. *Let Y_t be integrable and $\mathcal{F}_t$-measurable, and set $D_t = Y_t - \mathbb{E}(Y_t \mid \mathcal{F}_{t-1})$. If $\{|Y_t| : t \ge 1\}$ is uniformly integrable, then*

$$\frac{1}{N} \sum_{t=k+1}^{k+N} D_t \xrightarrow{p} 0$$

uniformly over deterministic starting indices $k \ge 0$ as $N \to \infty$.

Proof. For $M > 0$, let $Y_t^{(M)} = Y_t \mathbf{1}(|Y_t| \le M)$ and $R_t^{(M)} = Y_t - Y_t^{(M)}$. The bounded martingale differences $D_t^{(M)} = Y_t^{(M)} - \mathbb{E}(Y_t^{(M)} \mid \mathcal{F}_{t-1})$ satisfy

$$\mathbb{E}\left| \frac{1}{N} \sum_{t=k+1}^{k+N} D_t^{(M)} \right|^2 \le \frac{4M^2}{N}.$$

The remaining average has expected absolute value at most

$$\frac{2}{N}\sum_{t=k+1}^{k+N}\mathbb{E}|R_t^{(M)}| \le 2\sup_t \mathbb{E}\{|Y_t|\mathbf{1}(|Y_t|>M)\}.$$

First let $N\to\infty$ and then $M\to\infty$. The bounds do not depend on k, which proves the result. □

Proof of Theorem 3.6. The score convergence in Assumption 3.5 and the target second moment give $\sup_t \mathbb{E}\|G_t\|^2<\infty$. Cauchy–Schwarz yields

$$\mathbb{E}\|G_tG_t^{\mathsf{T}}-\xi_t^\star(\xi_t^\star)^{\mathsf{T}}\|_{\mathrm{F}} \le \big\{(\mathbb{E}\|G_t\|^2)^{1/2}+(\mathbb{E}\|\xi_t^\star\|^2)^{1/2}\big\}(\mathbb{E}\|G_t-\xi_t^\star\|^2)^{1/2}\longrightarrow 0.$$

The average of this bound over any deterministic retained interval with $N_n\to\infty$ tends to zero: a fixed finite prefix contributes $o(1)$, and every sufficiently late term is arbitrarily small. Markov's inequality reduces $\widehat{\mathrm{S}}_{n,b}$ to the average of the i.i.d. integrable target outer products, whose weak law gives S. Continuity at nonsingular A gives $\mathrm{R}_n\xrightarrow{p}(\mathrm{A}^{\mathsf{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathsf{T}}=\mathrm{A}^{-1}$, and another continuous mapping gives consistency.

Under the additional assumptions of Theorem 3.4, the contrast numerator has limiting variance $a^{\mathsf{T}}\Sigma a>0$. Slutsky's theorem gives the standardized normal limit and the coverage of (6). If Σ is positive definite, the corresponding quadratic form has the chi-squared limit, which gives the stated Wald ellipsoid. Consistency makes the estimated contrast variance positive, or the estimated covariance positive definite, with probability tending to one under the respective assumptions. Withholding the corresponding report on the complementary event thus leaves limiting coverage unchanged. □

C.7.5 Proofs for the model examples

Proof of corollary C.9. Take $g=\nabla R$, $\mathrm{A}=\mathrm{H}$, and $\mathrm{S}=\mathrm{V}_\ell$ in Theorem 3.4. The assumptions in the corollary are exactly Assumptions 3.1 and 3.2, while Assumptions 2.1 and 3.3 are imposed. This gives the normal limit stated in corollary C.9. □

Proof of corollary C.10. The mean score is exactly $g(\theta)=\mathrm{Q}(\theta-\theta_{\mathrm{LS}})$, so $\mathrm{A}=\mathrm{Q}$ and the mean remainder is zero. At the target, $\xi_t^\star=-X_t\varepsilon_t$, where $\varepsilon_t=Y_t-X_t^{\mathsf{T}}\theta_{\mathrm{LS}}$, and the target-score moment condition in corollary C.10 gives its finite second moment and covariance $\mathbb{E}(\varepsilon^2XX^{\mathsf{T}})$.

It remains to control the score variation; this step alone does not impose a fourth moment. Put

$$\mathrm{M}_t=X_tX_t^{\mathsf{T}}-\mathrm{Q},\qquad \Delta\xi_t=\mathrm{M}_te_{t-1},\qquad e_{t-1}=\theta_{t-1}-\theta_{\mathrm{LS}}.$$

The sequence $\Delta\xi_t$ is a martingale difference because e_{t-1} is $\mathcal{F}_{t-1}$-measurable and M_t is independent of $\mathcal{F}_{t-1}$ with mean zero. By lemma C.23, Lyapunov's inequality, and (10),

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{\sqrt{N_n}}\sum_{t=b_n+1}^{n}\Delta\xi_t\right\|^p &\le CN_n^{-p/2}\sum_{t=b_n+1}^{n}\mathbb{E}\|\mathrm{M}_te_{t-1}\|^p\\ &\le Cn^{-p/2}\sum_{t=1}^{n}t^{-\varrho p/2}.\end{aligned}$$

The second inequality uses independence of M_t and e_{t-1} together with

$$\mathbb{E}\|\mathrm{M}_t\|_{\mathrm{F}}^p<\infty,\qquad \mathbb{E}\|e_{t-1}\|^p\le\{\mathbb{E}\|e_{t-1}\|^2\}^{p/2}.$$

The last expression is bounded by

$$\begin{cases}O\{n^{1-p(1+\varrho)/2}\}, & \varrho p/2<1,\\ O(n^{-p/2}\log n), & \varrho p/2=1,\\ O(n^{-p/2}), & \varrho p/2>1.\end{cases}$$

Each case tends to zero under (21). Thus the score-variation term is $o_p(1)$.

All remaining terms in the proof of Theorem 3.4 use only Assumptions 2.1 and 3.3 and the zero mean remainder. Repeating those steps gives

$$\sqrt{N_n}(\bar{\theta}_{n,b} - \theta_{\mathrm{LS}}) = \mathrm{Q}^{-1} \frac{1}{\sqrt{N_n}} \sum_{t=b_n+1}^{n} X_t \varepsilon_t + o_p(1).$$

The i.i.d. central limit theorem proves corollary C.10. Under conditional homoskedasticity, $\mathbb{E}(\varepsilon^2 X X^\mathsf{T}) = \sigma^2 \mathrm{Q}$. □

Proof of corollary C.11. Let $q = X^\mathsf{T}\theta_{\log}$ and $a = X^\mathsf{T}e$. For deterministic $e \to 0$,

$$\frac{X\{\pi(q+a) - \pi(q) - \pi(q)\{1-\pi(q)\}a\}}{\|e\|} \longrightarrow 0 \quad \text{almost surely.}$$

Because $|\pi(q+a) - \pi(q)| \le |a|/4$ and $\pi(q)\{1-\pi(q)\} \le 1/4$, the norm of the displayed numerator divided by $\|e\|$ is bounded by $\|X\|^2/2$. Dominated convergence under $\mathbb{E}\|X\|^2 < \infty$ gives $g(\theta_{\log} + e) = \mathrm{A}_{\log} e + o(\|e\|)$.

For the sample score,

$$\|G(\theta_{\log} + e; X, Y) - G(\theta_{\log}; X, Y)\|^2 \le \|X\|^2,$$

and the left side tends to zero almost surely as $e \to 0$. Dominated convergence verifies local quadratic-mean continuity after centering. At the target the score covariance is $\mathrm{S}_{\log}$ in corollary C.11. Apply Theorem 3.4.

For the stated sufficient remainder bound, the logistic derivative is bounded and Lipschitz. Hence, for every $\nu \in (0, 1]$,

$$|\pi(q+a) - \pi(q) - \pi(q)\{1-\pi(q)\}a| \le C_\nu |a|^{1+\nu}.$$

It follows that $\|\mathfrak{r}(e)\| \le C\mathbb{E}\|X\|^{2+\nu}\|e\|^{1+\nu}$ whenever that moment is finite. Under correct Bernoulli specification, $\mathbb{E}[(Y - \pi_\star)^2 \mid X] = \pi_\star(1-\pi_\star)$, so $\mathrm{S}_{\log} = \mathrm{A}_{\log}$. □

Proof of corollary C.12. Write $\mathrm{H}(\theta) = \mathbb{E}[\pi_\theta(X)\{1-\pi_\theta(X)\}XX^\mathsf{T}]$. Positive definiteness of $\mathrm{A}_{\log}$ implies that $\mathrm{Q} = \mathbb{E}(XX^\mathsf{T})$ is positive definite. For every finite θ and nonzero z, the positive logistic weight and $\mathbb{P}(X^\mathsf{T}z \neq 0) > 0$ imply $z^\mathsf{T}\mathrm{H}(\theta)z > 0$. The bound $\pi_\theta(1-\pi_\theta)\|X\|^2 \le \|X\|^2/4$ shows that H is continuous. Compactness gives $m_{\mathcal{C}} = \inf_{\theta\in\mathcal{C}} \lambda_{\min}\{\mathrm{H}(\theta)\} > 0$. For $e = \theta - \theta_{\log}$ with $\theta \in \mathcal{C}$, convexity of $\mathcal{C}$ and integration along the segment imply

$$e^\mathsf{T} g(\theta) = \int_0^1 e^\mathsf{T}\mathrm{H}(\theta_{\log} + se)e \, ds \ge m_{\mathcal{C}}\|e\|^2.$$

The derivative of the scalar logistic function is bounded and Lipschitz, hence globally ν-Hölder for $0 < \nu \le 1$. Its integrated remainder, dominated by $C\|X\|^{2+\nu}\|e\|^{1+\nu}$, gives $\|\mathfrak{r}(e)\| \le C\|e\|^{1+\nu}$. The growth bound follows from $\|\mathrm{H}(\theta)\| \le \mathbb{E}\|X\|^2/4$; the score and centered-noise second moments are uniformly bounded by multiples of $\mathbb{E}\|X\|^2$. Since $\mathrm{K} = cp_0\mathrm{I}_d/(1-\beta)$, these bounds verify Theorem C.6 on $\mathcal{C}$. The mean/score continuity and natural-Jacobian conclusions follow as in corollary C.11 and proposition C.15. No smallness condition on $\mathcal{C}$ was used. □

Proof of corollary C.14. Conditioning on X and integrating the derivative of the clipped score gives

$$\begin{aligned} g_H(\theta_H + e) - g_H(\theta_H) &= \mathbb{E}\left[X \int_0^{X^\mathsf{T}e} p_X(s)\, ds\right] \\ &= \mathrm{A}_H e + \mathbb{E}\left[X \int_0^{X^\mathsf{T}e} \{p_X(s) - p_X(0)\}\, ds\right], \end{aligned}$$

where the second term is $o(\|e\|)$ by (22).

For each fixed (X, Y), continuity of ψ_{c_H} gives $G_H(\theta_H + e; X, Y) - G_H(\theta_H; X, Y) \to 0$. Its squared norm is bounded by $4c_H^2\|X\|^2$, so dominated convergence verifies local quadratic-mean continuity after centering. At the target, $G_H(\theta_H; X, Y) = -X\psi_{c_H}(\varepsilon_H)$, which has covariance S_H. Apply Theorem 3.4. □

Proof of proposition C.15. The least-squares statement is the ordinary weak law of large numbers. For logistic regression, let $\mathrm{H}_t = \pi_{\theta_{t-1}}(X_t)\{1 - \pi_{\theta_{t-1}}(X_t)\}X_t X_t^\mathsf{T}$. Each entry is dominated in absolute value by $\|X_t\|^2/4$, so the entries are uniformly integrable. The martingale weak law in lemma C.28 gives $n^{-1}\sum_t\{\mathrm{H}_t - \mathbb{E}(\mathrm{H}_t \mid \mathcal{F}_{t-1})\} = o_p(1)$. The conditional mean converges to $\mathrm{A}_{\log}$ by dominated convergence and $\theta_{t-1} \xrightarrow{p} \theta_{\log}$; bounded convergence in probability then gives the Cesàro result.

For Huber regression, set $\mathrm{H}_t = X_t X_t^\mathsf{T}\mathbf{1}\{|Y_t - X_t^\mathsf{T}\theta_{t-1}| < c_H\}$. The entries are dominated by $\|X_t\|^2$ and are therefore uniformly integrable. The martingale part is again $o_p(1)$. Conditional on a deterministic $e \to 0$, the indicator converges almost surely to $\mathbf{1}(|\varepsilon_H| < c_H)$ where $X \neq 0$ under the weighted no-boundary-mass condition. Where $X = 0$, the matrix integrand vanishes regardless of the indicator. Dominated convergence and $e_{t-1} \xrightarrow{p} 0$ show that the conditional mean converges in probability and in L^1 to A_H. Cesàro averaging proves $\widehat{\mathrm{A}}_{H,n} \xrightarrow{p} \mathrm{A}_H$.

Under Assumption C.13, the weighted boundary condition already follows from the specified derivative. For every vector z, dominated convergence in the symmetric directional difference quotient gives

$$\lim_{h\downarrow 0}\frac{z^\mathsf{T}\{g_H(\theta_H + hz) - g_H(\theta_H - hz)\}}{2h} = z^\mathsf{T}\mathrm{A}_H z + \tfrac{1}{2}\mathbb{E}[(X^\mathsf{T}z)^2\mathbf{1}\{|\varepsilon_H| = c_H\}].$$

The left side is $z^\mathsf{T}\mathrm{A}_H z$ by Assumption C.13. Taking coordinate directions proves the weighted boundary condition.

For logistic regression, $\|G(\theta; X, Y)\| \leq \|X\|$; for Huber regression, $\|G_H(\theta; X, Y)\| \leq c_H\|X\|$. Thus $\mathbb{E}\|X\|^2 < \infty$ gives uniform integrability of squared scores. Their local L^2 convergence was established in the proofs of corollaries C.11 and C.14. Applying Theorem 3.6 completes the statements. □

C.7.6 Proofs for quantile regression

Proof of proposition C.16. Write $q_\tau(X) = X^\mathsf{T}\theta_\tau$ and $a = X^\mathsf{T}e$. Conditioning on X and using $g_\tau(\theta_\tau) = 0$,

$$\begin{aligned} g_\tau(\theta_\tau + e) - g_\tau(\theta_\tau) &= \mathbb{E}\left[X\{F(q_\tau(X) + a \mid X) - F(q_\tau(X) \mid X)\}\right] \\ &= \mathrm{A}_\tau e + \mathfrak{r}_\tau(e), \end{aligned}$$

where $\mathfrak{r}_\tau(e)$ is the vector inside the norm in (15). Hence $\|\mathfrak{r}_\tau(e)\| = o(\|e\|)$, proving the stated mean linearization.

Let $\Delta G = X\{\mathbf{1}(Y \leq q_\tau(X) + a) - \mathbf{1}(Y \leq q_\tau(X))\}$. For almost every (X, Y) with $Y \neq q_\tau(X)$, the indicator difference tends to zero as $e \to 0$. The no-atom condition in Assumption 3.9 makes this an almost-sure statement, and $\|\Delta G\|^2 \leq \|X\|^2$. Dominated convergence under $\mathbb{E}\|X\|^2 < \infty$ proves the claimed quadratic-mean score continuity. Since centering cannot increase the second moment, the same conclusion holds for the centered score. Defining the modulus as the supremum over a sufficiently small parameter ball verifies Assumption 3.2.

□

Proof of Theorem C.17. By proposition C.16, the mean and score conditions of Theorem 3.4 hold with $\mathrm{A} = \mathrm{A}_\tau$. Also $\|G_\tau(\theta_\tau; X, Y)\| \leq \|X\|$, so the target second moment is finite. Applying Theorem 3.4 gives the stated normal limit. Under correct conditional quantile specification, the indicator is conditionally Bernoulli with success probability τ, yielding $\mathrm{S}_\tau = \tau(1-\tau)\mathbb{E}(XX^\mathsf{T})$. Under the additional stated Bahadur-expansion condition, the batch-equivalence statement follows from corollary C.3. □

Proof of Theorem 3.11. Define

$$\mathrm{H}_t = \frac{1}{h_t}\mathcal{K}\left(\frac{Y_t - X_t^\mathsf{T}\theta_{t-1}}{h_t}\right)X_t X_t^\mathsf{T}.$$

Decompose

$$\widehat{\mathrm{A}}_{\tau,n} - \mathrm{A}_\tau = \frac{1}{n}\sum_{t=1}^{n}\{\mathrm{H}_t - \mathbb{E}(\mathrm{H}_t \mid \mathcal{F}_{t-1})\} + \frac{1}{n}\sum_{t=1}^{n}\{\mathbb{E}(\mathrm{H}_t \mid \mathcal{F}_{t-1}) - \mathrm{A}_\tau\}. \tag{55}$$

Step 1: martingale term. For the $p \in (1, 2]$ in Assumption 3.10, conditional integration gives

$$\mathbb{E}\{\|\mathrm{H}_t\|_{\mathrm{F}}^p \mid X_t, \mathcal{F}_{t-1}\} \le h_t^{1-p}\bar{f}(X_t)\,\|X_t\|^{2p}\int \mathcal{K}(u)^p\,du.$$

Thus $\mathbb{E}\|\mathrm{H}_t\|_{\mathrm{F}}^p \le Ch_t^{1-p}$. Conditional Jensen's inequality shows that the centered term $\mathrm{H}_t - \mathbb{E}(\mathrm{H}_t \mid \mathcal{F}_{t-1})$ has the same order of pth moment. Applying lemma C.23 to the vectorized matrix martingale,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\{\mathrm{H}_t - \mathbb{E}(\mathrm{H}_t \mid \mathcal{F}_{t-1})\}\right\|_{\mathrm{F}}^p &\le Cn^{-p}\sum_{t=1}^{n}h_t^{1-p}\\ &= O\{n^{-(p-1)(1-\kappa)}\} = o(1).\end{aligned}$$

Step 2: conditional-mean bias. Let $e_{t-1} = \theta_{t-1} - \theta_\tau$. Let $\mathbb{E}_X$ denote expectation over the marginal law of a fresh regressor X, holding e_{t-1} fixed. A change of variables gives

$$\mathbb{E}(\mathrm{H}_t \mid \mathcal{F}_{t-1}) = \mathbb{E}_X\left[XX^{\mathsf{T}}\int \mathcal{K}(u)f(q_\tau(X) + X^{\mathsf{T}}e_{t-1} + h_t u \mid X)\,du\right].$$

By Assumption 3.10, compact support of $\mathcal{K}$, and $|a+b|^{\nu_A} \le |a|^{\nu_A} + |b|^{\nu_A}$,

$$\begin{aligned}\|\mathbb{E}(\mathrm{H}_t \mid \mathcal{F}_{t-1}) - \mathrm{A}_\tau\|_{\mathrm{F}} \le C\,\|e_{t-1}\|^{\nu_A}\,\mathbb{E}\{L_A(X)\,\|X\|^{2+\nu_A}\}\\ + Ch_t^{\nu_A}\mathbb{E}\{L_A(X)\,\|X\|^2\}.\end{aligned}$$

The assumed Cesàro displacement-moment condition and $h_t \to 0$ control the bias after m. The conditional means in the finite prefix are integrable by Step 1, so their average vanishes in L^1. Thus the second term in (55) converges to zero in L^1. Hence $\widehat{\mathrm{A}}_{\tau,n} \xrightarrow{p} \mathrm{A}_\tau$ and $\widetilde{\mathrm{A}}_{\tau,n}^{-1} \xrightarrow{p} \mathrm{A}_\tau^{-1}$.

Step 3: score covariance. The bound $\|G_\tau(\theta; X, Y)\| \le \|X\|$ and $\mathbb{E}\|X\|^2 < \infty$ imply uniform integrability of the squared online scores. The map

$$e \longmapsto \mathbb{E}\,\|G_\tau(\theta_\tau + e; Z) - G_\tau(\theta_\tau; Z)\|^2$$

is bounded and continuous at zero by proposition C.16. Since $e_{t-1} \to 0$ in probability, bounded convergence in probability gives

$$\mathbb{E}\,\|G_\tau(\theta_{t-1}; Z) - G_\tau(\theta_\tau; Z)\|^2 \longrightarrow 0.$$

The score part of Assumption 3.5 therefore holds. Together with the Jacobian consistency proved above, Theorem 3.6 gives $\widehat{\mathrm{S}}_{\tau,n,b} \xrightarrow{p} \mathrm{S}_\tau$. Continuous mapping proves Theorem 3.11. □