STOCHASTIC NEWTON WITHOUT HESSIANS: ESTIMATION AND INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic Newton methods use population curvature, which may remain well defined when sample Hessians are unavailable or undefined. We develop an online stochastic Newton method based on symmetric score secants and Polyak-Ruppert averaging that allows discontinuous sample scores. A spectrally projected recursive average supplies the curvature matrix for subsequent parameter updates. Under stability, iterate rate and score regularity conditions, we derive restrictions on secant second moments, the perturbation radius and the auxiliary batch size under which curvature learning is asymptotically negligible in the Polyak-Ruppert expansion. For quantile regression, shared endpoint observations give an $O(\delta^{-1})$ secant second moment bound, whereas independent endpoints have variance of order δ^{-2} under nondegeneracy, so sharing enlarges the sufficient tuning region for the curvature radius. For direct unbiased primary scores, the averaged estimator retains the usual first-order sandwich limit, and when the spectral projection set contains the population curvature the learned matrix also supports consistent online sandwich estimation. We separately analyze the smoothing bias and direction-dependent covariance of function-value primary updates with auxiliary score access retained, and under additional process conditions a centered functional limit yields iterate-based self-normalized inference. Simulations examine the predicted rates and finite-sample behavior, and an online news application illustrates streaming quantile inference.

1 INTRODUCTION

Streaming data call for parameter estimates and uncertainty measures that can be updated as observations arrive, which stochastic approximation provides through recursive score based updates. In fixed dimension, Polyak-Ruppert averaging yields an asymptotically linear estimator under suitable regularity and stepsize conditions (Ruppert, 1988; Polyak & Juditsky, 1992; Kushner & Yin, 2003), providing a basis for inference without repeatedly fitting the model to all past observations (Chen et al., 2020; Zhu et al., 2023).

Stochastic Newton methods update an estimate of population curvature along with the parameter. Existing stochastic Newton and conditional gradient procedures typically obtain it from sample Hessians or smooth curvature approximations (Bercu et al., 2020; Leluc & Portier, 2023; Boyer & Godichon-Baggioni, 2023; Godichon-Baggioni & Werge, 2025), which are not always available. In quantile regression, for example, the sample loss is piecewise linear with a zero Hessian wherever it exists, whereas the population risk can have a positive definite Hessian (Koenker, 2005); the primary update may also be restricted to function evaluations when sample gradients are inaccessible. The problem is then to carry out Newton estimation with limited derivative information and provide valid inference for the resulting Polyak-Ruppert averaging estimator.

We propose a stochastic Newton method based on score secants and establish conditions under which curvature learning preserves the first order distribution of the Polyak-Ruppert averaging estimator. For a differentiable vector field $g : \mathbb{R}^d \to \mathbb{R}^d$ with Jacobian $Dg(\theta)$, $[Dg(\theta)]_{ij} = \partial g_i(\theta)/\partial\theta_j$, the target is defined by $g(\theta_\star) = 0$, where g has a continuous symmetric Jacobian near $\theta_\star$ and $\mathrm{A} = Dg(\theta_\star)$ is positive definite. The stochastic Newton direction $s_t \in \mathbb{R}^d$ is the unique solution of the linear system in the update $\mathrm{B}_{t-1}s_t = \widehat{G}_t(\theta_{t-1})$, $\theta_t = \theta_{t-1} - \eta_t s_t$, where $\widehat{G}_t$ is the primary stochastic score, B_{t-1} is a predictable positive definite curvature estimate, and $\eta_t \asymp t^{-\alpha}$ with $1/2 < \alpha < 1$. When B_t

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.

consistently estimates A, the local mean update matrix $\mathrm{B}_{t-1}^{-1}\mathrm{A}$ converges to I_d, the oracle Newton scaling. Inference uses the Polyak-Ruppert average $\bar{\theta}_n = n^{-1}\sum_{t=1}^{n}\theta_{t-1}$ and the curvature matrix of the recursion.

Curvature estimation needs only paired evaluations of an auxiliary score Ψ with $\mathbb{E}\{\Psi(\theta; W)\} = g(\theta)$. For a bounded random direction V independent of W with $\mathbb{E}(VV^{\mathsf{T}}) = \mathrm{I}_d$, define $\mathrm{K}_\delta(\theta; W, V) = \mathrm{sym}[\{\Psi(\theta + \delta V; W) - \Psi(\theta - \delta V; W)\}V^{\mathsf{T}}/(2\delta)]$, where $\mathrm{sym}(\mathrm{C}) = (\mathrm{C} + \mathrm{C}^{\mathsf{T}})/2$. Its expectation approximates the population Jacobian without differentiating the sample score; B_t recursively averages auxiliary batches of these matrices under a spectral projection that maintains positive definiteness, and each secant uses the same observation W at both endpoints.

Stochastic curvature approximations and common random number finite differences are well established (Spall, 2000; Mokhtari & Ribeiro, 2014; Li et al., 2018; Zhu et al., 2020; Kleinman et al., 1999). For online inference, the central issue is that shrinking the perturbation radius improves the curvature approximation but can increase secant variability, and the changing Newton matrix then adds a remainder to the averaged estimator. An & Huo (2026) identify a sufficient stabilization threshold for this remainder; we connect it to score regularity, perturbation radius and auxiliary batch size through bounds on the L^2 increments of the curvature recursion. Our contributions are as follows.

1. **Curvature learning without sample Hessians.** Paired score evaluations estimate the population Jacobian even for discontinuous sample scores, and their second moments determine the fluctuations of the recursively learned matrix. For quantile scores, common observations give exponent $q = 1$ versus $q = 2$ for independent endpoints under nondegeneracy, enlarging the sufficient radius/batch region.
2. **Inference with learned curvature.** An asymptotic linear representation accounts for the changing Newton matrix and primary score approximation bias; under stability and regularity assumptions, direct unbiased scores retain the averaged oracle Newton limit, and recursively averaged score outer products with consistent learned curvature yield an online sandwich estimator, also under misspecification.
3. **Model specializations and extensions.** We verify the score conditions for least squares, logistic, Huber and quantile regression with the iterate rate inputs explicit, separate primary finite difference bias from curvature learning while retaining auxiliary score access, and give self-normalized inference under a centered functional limit.

1.1 Related work

Online uncertainty quantification. The averaging theory of Ruppert (1988) and Polyak & Juditsky (1992) underlies inference from stochastic gradient iterates. Chen et al. (2020) develop plug-in and batch-means estimators of the asymptotic covariance of averaged SGD, and Zhu et al. (2023) recursive estimators that need no prespecified terminal sample size; the plug-in approach estimates sensitivity and score covariance separately, retaining the sandwich structure needed under misspecification (White, 1982). The online bootstrap of Fang et al. (2018) instead runs randomly perturbed SGD recursions alongside the original one, with quantile regression among its applications. Our sandwich construction uses the curvature matrix already maintained for the Newton updates, together with recursively averaged score outer products.

Self-normalization is another route: Shao (2010a;b) replace the long run variance estimator by a random normalizer from recursive estimates, and Lee et al. (2022) apply this to averaged SGD through a functional central limit theorem. Our functional extension uses the same principle; it additionally requires uniform control of the remainder from the changing Newton matrix, while the pivotal random scaling distribution is inherited from existing theory.

Stochastic Newton methods and curvature estimation. Stochastic Newton methods adapt recursive root finding (Robbins & Monro, 1951; Ruppert, 1985) to the derivative of the mean score; later work gives limiting distributions for logistic regression (Bercu et al., 2020), convergence rates for weighted averages (Boyer & Godichon-Baggioni, 2023; 2024), the effect of conditioning (Leluc & Portier, 2023), expected cost of order dN under randomized matrix updates and batching (Godichon-Baggioni & Werge, 2025), and last-iterate inference with Hessian averaging and randomized linear

solves (Wang et al., 2026). Building on the stabilization analysis of An & Huo (2026), we instead relate changes in a recursively averaged secant matrix to score second moments and endpoint coupling, including discontinuous scores.

Curvature estimation from evaluations is also well established: Spall (2000) develops second order stochastic approximation from simultaneous perturbations of function or gradient measurements, Li et al. (2018) approximate Hessian vector products by stochastic gradient differences in a Newton-based inference procedure for smooth empirical risks, and Mokhtari & Ribeiro (2014); Zhu et al. (2020) study related stochastic quasi-Newton implementations. These methods show that sample Hessians are not essential to second-order updating; our analysis concerns a recursively averaged score secant matrix and how nonsmooth sample scores affect its fluctuations and the distribution of averaged Newton iterates.

Nonsmooth stochastic approximation and online quantile regression. Davis et al. (2024) establish asymptotic normality and optimality for nonsmooth stochastic approximation under structural assumptions yielding a local smooth representation, and Jiang et al. (2025) bound the expected operator norm error of online batch means in nonsmooth, possibly nonmonotone stochastic approximation. In quantile regression, Shen et al. (2025) study nonasymptotic estimation error, regret and initialization, whereas Lee et al. (2025) establish an averaged limit with the batch estimator's first-order covariance and develop random scaling inference. These results are distributional and computational benchmarks for the present quantile specialization, which additionally estimates and uses population curvature from discontinuous scores under the stated stability and iterate rate assumptions. The shared observation secant is related to common random numbers in finite difference stochastic approximation (Kleinman et al., 1999); here the comparison with independent endpoints determines different sufficient tuning regions for curvature learning.

Smoothing and function-value primary updates. Fernandes et al. (2021) smooth the quantile regression objective and analyze the resulting bias and inference; our direct score procedure leaves the primary quantile score unsmoothed and perturbs only the auxiliary score. When only sample loss evaluations are available, the Kiefer-Wolfowitz construction (Kiefer & Wolfowitz, 1952) approximates the primary score by finite differences of function values, and Chen et al. (2024) develop online inference for such procedures, characterizing how the limiting covariance varies with the number and dependence of random directions. Our finite difference primary extension uses this covariance calculation; the primary smoothing radius controls accumulated score bias, the auxiliary radius and batch size control curvature stabilization, and the auxiliary update still requires score evaluations.

Organization. Sections 2 to 4 present the method, its asymptotic guarantees and numerical evidence; the appendices give simulation details (Section A), the news application (Section B), model-specific results, self-normalized inference and proofs (Section C).

Notation. $\|\cdot\|$ is the Euclidean norm of a vector, $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{F}}$ the operator and Frobenius norms of a matrix, I_d the $d \times d$ identity and a superscript T transposition. For symmetric matrices, $\mathrm{C} \preceq \mathrm{D}$ means that $\mathrm{D} - \mathrm{C}$ is positive semidefinite; $\xrightarrow{p}$ and $\Longrightarrow$ denote convergence in probability and in distribution, and constants C may change between displays.

2 METHODOLOGY

Let $\mathrm{H}(\theta) = Dg(\theta)$ and $\mathrm{A} = \mathrm{H}(\theta_\star)$ in fixed dimension; H is continuous and symmetric near the root with A positive definite, while the sample score may be discontinuous. With $\mathcal{F}_t$ the initialization, observations and directions through iteration t, write for the primary score

$$\mathbb{E}\{\widehat{G}_t(\theta) \mid \mathcal{F}_{t-1}\} = g(\theta) + b_t(\theta), \tag{1}$$

$$\zeta_t(\theta) = \widehat{G}_t(\theta) - \mathbb{E}\{\widehat{G}_t(\theta) \mid \mathcal{F}_{t-1}\}. \tag{2}$$

Here $\zeta_t(\theta)$ is the centered fluctuation of the primary score. A fresh direct score $G(\theta; Z_t)$ with mean $g(\theta)$ has $b_t = 0$; the appendix also treats function-value primary scores, whose approximation bias is nonzero, with auxiliary score access retained.

Let the auxiliary score satisfy

$$\mathbb{E}\{\Psi(\theta; W)\} = g(\theta). \tag{3}$$

To estimate the Jacobian without sample Hessians, form a score secant along a random direction V_t, which estimates the Jacobian-vector product $\mathrm{H}(\theta_{t-1})V_t$:

$$D_{tj} = \frac{\Psi(\theta_{t-1} + \delta_t V_t; W_{tj}) - \Psi(\theta_{t-1} - \delta_t V_t; W_{tj})}{2\delta_t}, \tag{4}$$

$$\widehat{\mathrm{H}}_t = R_t^{-1} \sum_{j=1}^{R_t} \mathrm{sym}(D_{tj} V_t^{\mathsf{T}}). \tag{5}$$

The conditional mean of the noisy secant $D_{tj} \in \mathbb{R}^d$ approaches $\mathrm{H}(\theta_{t-1})V_t$; since $\mathbb{E}(V_t V_t^{\mathsf{T}}) = \mathrm{I}_d$ and $\mathrm{H}(\theta)$ is symmetric, averaging $\mathrm{sym}(D_{tj} V_t^{\mathsf{T}})$ recovers the population Jacobian. All observations in the auxiliary batch share the direction V_t.

For fixed $0 < \ell < u$, let $\mathcal{S}_{\ell,u} = \{\mathrm{B} = \mathrm{B}^{\mathsf{T}} : \ell \mathrm{I}_d \preceq \mathrm{B} \preceq u \mathrm{I}_d\}$ with Frobenius projection $\Pi_{\ell,u}$, equivalently eigenvalue clipping. Initialize $\mathrm{B}_0 \in \mathcal{S}_{\ell,u}$, θ_0, $\bar{\theta}_0 = 0$, $\widehat{\mathrm{Q}}_0 = 0$ and, with $\eta_t = \eta_0 t^{-\alpha}$, $1/2 < \alpha < 1$, perform

$$\theta_t = \theta_{t-1} - \eta_t \mathrm{P}_t \widehat{G}_t(\theta_{t-1}), \qquad \mathrm{P}_t = \mathrm{B}_{t-1}^{-1}, \tag{6}$$

$$\mathrm{B}_t = \Pi_{\ell,u}\{(1 - t^{-1})\mathrm{B}_{t-1} + t^{-1}\widehat{\mathrm{H}}_t\}. \tag{7}$$

The update at t solves with B_{t-1}; the new matrix B_t is first used at $t+1$. The Polyak-Ruppert average and the score second moment matrix are updated recursively as

$$\bar{\theta}_t = (1 - t^{-1})\bar{\theta}_{t-1} + t^{-1}\theta_{t-1} = t^{-1} \sum_{s=1}^{t} \theta_{s-1}, \tag{8}$$

$$\widehat{\mathrm{Q}}_t = (1 - t^{-1})\widehat{\mathrm{Q}}_{t-1} + t^{-1} Y_t Y_t^{\mathsf{T}}, \qquad Y_t = \widehat{G}_t(\theta_{t-1}), \tag{9}$$

$$\widehat{\Sigma}_t = \mathrm{B}_t^{-1} \widehat{\mathrm{Q}}_t \mathrm{B}_t^{-1}. \tag{10}$$

Under the centered transfer and covariance conditions below, a contrast $c^{\mathsf{T}}\theta_\star$ with positive limiting variance has the asymptotic interval

$$c^{\mathsf{T}}\bar{\theta}_t \;\pm\; z_{1-\gamma/2}\sqrt{c^{\mathsf{T}}\widehat{\Sigma}_t c / t}. \tag{11}$$

Here z_a is the standard normal quantile and t counts primary updates. Figure 6 in the appendix summarizes the procedure, and Proposition C.8 gives conditions for reusing the current primary observation in the auxiliary batch, which preserves predictability only when the new curvature is used at the next iteration.

An individual symmetrized secant need not be positive definite, even when the population curvature is. Projection therefore regularizes the matrix before it is used in a linear solve: every eigenvalue of B_t^{-1} lies in $[u^{-1}, \ell^{-1}]$. This ensures that each step is computable, but the iterate rate still depends on the population drift and noise conditions stated below. For a single contrast, solve $\mathrm{B}_t v = c$ and evaluate $v^{\mathsf{T}}\widehat{\mathrm{Q}}_t v / t$ to obtain the variance in (11), without forming the full matrix.

3 Theoretical results

Let $e_t = \theta_t - \theta_\star$ and assume that

$$g(\theta_\star + e) = \mathrm{A}e + r(e), \tag{12}$$

with A nonsingular; the transfer theorem does not need the symmetry that the score secant specialization adds. Define $\mathrm{M}_t = (\mathrm{P}_t \mathrm{A})^{-1} = \mathrm{A}^{-1}\mathrm{P}_t^{-1}$.

Assumption 3.1. *The following conditions hold.*

(T1) ***Iterate rate.*** *The recursion is well defined,* $\mathbb{E}\,\|e_0\|^2 < \infty$*, and*

$$\mathbb{E}\,\|e_t\|^2 \le C\eta_t. \tag{13}$$

(T2) ***Dynamic stabilization.*** *The effective inverse drifts are uniformly bounded in L^2 and, for some $\beta_{\mathrm{P}} > (1+\alpha)/2$,*

$$\sup_{t\geq 1}\{\mathbb{E}\,\|\mathrm{M}_t\|_{\mathrm{op}}^2\}^{1/2} < \infty, \qquad \{\mathbb{E}\,\|\mathrm{M}_t - \mathrm{M}_{t-1}\|_{\mathrm{op}}^2\}^{1/2} \leq Ct^{-\beta_{\mathrm{P}}} \quad (t \geq 2). \tag{14}$$

(T3) ***Mean remainder.***

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbb{E}\,\|r(e_{t-1})\| \longrightarrow 0. \tag{15}$$

(T4) ***Local score replacement.***

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\,\|\zeta_t(\theta_{t-1}) - \zeta_t(\theta_\star)\|^2 \longrightarrow 0. \tag{16}$$

(T5) ***Local oracle bias.***

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbb{E}\,\|b_t(\theta_{t-1}) - b_t(\theta_\star)\| \longrightarrow 0. \tag{17}$$

(T6) ***Target martingale limit.*** *With $\zeta_t^\star = \zeta_t(\theta_\star)$, for a deterministic positive semidefinite matrix* Q*,*

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}(\zeta_t^\star\zeta_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1}) \xrightarrow{p} \mathrm{Q}, \tag{18}$$

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Big[\|\zeta_t^\star\|^2\,\mathbf{1}\{\|\zeta_t^\star\| > \varepsilon\sqrt{n}\} \mid \mathcal{F}_{t-1}\Big] \xrightarrow{p} 0 \quad \textit{for every } \varepsilon > 0. \tag{19}$$

The rate in (T1) is a model-specific stability input. The identity $\mathrm{M}_t\mathrm{P}_t = \mathrm{A}^{-1}$ removes preconditioning from the leading score, and (T2) controls its Abel remainder; the decomposition and stabilization threshold follow An & Huo (2026). In the appendix, Proposition C.3 gives primitive pointwise sufficient conditions for (T3) to (T5), Proposition C.42 a risk-descent route to the iterate rate in (T1), and Section C.5 all proofs.

Theorem 3.2 (Transfer theorem for dynamic preconditioning)**.** *Suppose Assumption 3.1 holds and*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}b_t(\theta_\star) \xrightarrow{p} b \in \mathbb{R}^d. \tag{20}$$

Then

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\zeta_t^\star - \mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}b_t(\theta_\star) + o_p(1), \tag{21}$$

and consequently

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N\big(-\mathrm{A}^{-1}b,\ \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}\big). \tag{22}$$

Thus curvature learning preserves the averaged first-order covariance whenever its changing matrix leaves a negligible remainder: it affects finite-sample conditioning, not the oracle covariance.

The stabilization threshold follows from multiplying each matrix increment by the current estimation error. In the summation by parts decomposition, this contribution is bounded in mean by $Cn^{-1/2}\sum_{t=1}^{n-1}t^{\alpha/2-\beta_{\mathrm{P}}}$: the inverse stepsize contributes t^{α}, the matrix increment contributes $t^{-\beta_{\mathrm{P}}}$, and the iterate error contributes $t^{-\alpha/2}$. Thus (T2) makes this term vanish, while the endpoint and stepsize-change terms vanish because $\alpha < 1$; Section C.5.1 gives the full bound. This explains why consistency of the curvature estimate alone does not supply the transfer argument.

Curvature learning and endpoint coupling. The exponent q measures the cost of shrinking the secant radius: the population continuity, bounded isotropic direction and paired sample conditions in Assumption C.4 give a conditional variance bound of order $1 + (R_t\delta_t^q)^{-1}$ (Proposition C.5), whose constant term reflects the shared batch direction. Set

$$\delta_t = \delta_0 t^{-\kappa_{\mathrm{H}}}, \qquad R_t = \lceil R_0 t^r \rceil, \qquad \delta_0, R_0, \kappa_{\mathrm{H}} > 0, \quad r \geq 0, \qquad s_{\mathrm{H}} = (q\kappa_{\mathrm{H}} - r)_+. \tag{23}$$

Theorem 3.3. *Let* A *be symmetric positive definite and define* $\mathrm{A}_{\ell,u} := \Pi_{\ell,u}(\mathrm{A})$. *If* $\mathrm{A} = \mathrm{O}\,\mathrm{diag}(a_1, \ldots, a_d)\mathrm{O}^{\mathsf{T}}$, *then* $\mathrm{A}_{\ell,u} = \mathrm{O}\,\mathrm{diag}\{(a_j \vee \ell) \wedge u : 1 \leq j \leq d\}\mathrm{O}^{\mathsf{T}}$. *Suppose that, after deleting a finite prefix,*

$$d_{t,2} := \mathbb{E}\left\|\mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1}) - \mathrm{A}\right\|_{\mathrm{F}}^2 \longrightarrow 0, \tag{24}$$

$$\mathbb{E}\left[\left\|\widehat{\mathrm{H}}_t - \mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1})\right\|_{\mathrm{F}}^2 \mid \mathcal{F}_{t-1}\right] \leq C\left(1 + \frac{1}{R_t\delta_t^q}\right). \tag{25}$$

Let (7) *use* (23). *If* $s_{\mathrm{H}} < 1$, *then*

$$\mathbb{E}\,\|\mathrm{B}_t - \mathrm{A}_{\ell,u}\|_{\mathrm{F}}^2 \longrightarrow 0. \tag{26}$$

Moreover,

$$\{\mathbb{E}\,\|\mathrm{B}_t - \mathrm{B}_{t-1}\|_{\mathrm{F}}^2\}^{1/2} \leq Ct^{-1+s_{\mathrm{H}}/2}. \tag{27}$$

Consequently, with $\mathrm{P}_t = \mathrm{B}_{t-1}^{-1}$,

$$s_{\mathrm{H}} < 1 - \alpha \quad \textit{implies (14) with} \quad \beta_{\mathrm{P}} = 1 - \tfrac{s_{\mathrm{H}}}{2} > \tfrac{1+\alpha}{2}, \tag{28}$$

and $\mathrm{B}_t \xrightarrow{p} \mathrm{A}_{\ell,u}$, $\mathrm{P}_t \xrightarrow{p} \mathrm{A}_{\ell,u}^{-1}$, *so* $\mathrm{B}_t \xrightarrow{p} \mathrm{A}$ *if* $\mathrm{A} \in \mathcal{S}_{\ell,u}$. *Under Assumption C.4,* (24) *holds if* $\theta_t \xrightarrow{p} \theta_\star$ *and* $\theta_t \in \mathcal{U}$ *almost surely for all* $t \geq t_0$, t_0 *deterministic; a random entrance time needs a stopped argument or an integrable global domination condition for the same unconditional* L^2 *conclusion.*

Pathwise Lipschitz scores have $q = 0$ (Proposition C.6). For the quantile score G_τ in (29) below, sharing the observation makes a score difference nonzero only when the perturbation crosses a residual boundary; a bounded conditional density gives crossing probability $O(\delta)$, hence second moment $O(\delta^{-1})$, whereas independent endpoints retain noise as $\delta \to 0$ and have order δ^{-2} under nondegeneracy (Propositions C.13 and C.14).

Let $Z = (X, Y)$ and fix $\tau \in (0, 1)$; see Koenker (2005) for the standard framework. Define

$$G_\tau(\theta; X, Y) = X\{\mathbf{1}(Y \leq X^{\mathsf{T}}\theta) - \tau\}, \tag{29}$$

and let θ_τ be a locally unique solution of

$$g_\tau(\theta_\tau) = \mathbb{E}G_\tau(\theta_\tau; X, Y) = 0. \tag{30}$$

This definition permits misspecification; the stronger condition

$$\mathbb{P}(Y \leq X^{\mathsf{T}}\theta_\tau \mid X) = \tau \quad \text{a.s.} \tag{31}$$

is used only to simplify the covariance. The population score derivative is

$$\mathrm{A}_\tau = \mathbb{E}\{f_{Y|X}(X^{\mathsf{T}}\theta_\tau \mid X)XX^{\mathsf{T}}\}. \tag{32}$$

Assumption 3.4. *There are a convex neighborhood* $\mathcal{U}$ *of* θ_τ, *an open convex tube containing* $\mathcal{U}$ *and all its permitted secant perturbations, a bounded isotropic direction* V, *random envelopes* $\bar{f}(X)$ *and* $L_f(X)$, $\nu \in (0, 1]$, *and* $\epsilon > 0$ *such that:*

(Q1) for every θ_1, θ_2 *in this tube,*

$$f_{Y|X}(X^{\mathsf{T}}\theta_1 \mid X) \leq \bar{f}(X),$$
$$\left|f_{Y|X}(X^{\mathsf{T}}\theta_1 \mid X) - f_{Y|X}(X^{\mathsf{T}}\theta_2 \mid X)\right| \leq L_f(X)\left|X^{\mathsf{T}}(\theta_1 - \theta_2)\right|^\nu; \tag{33}$$

(Q2)

$$\mathbb{E}\,\|X\|^{2+\epsilon} < \infty, \qquad \mathbb{E}\{\bar{f}(X)(\|X\|^2 + \|X\|^3)\} < \infty, \qquad \mathbb{E}\{L_f(X)\,\|X\|^{2+\nu}\} < \infty; \tag{34}$$

(Q3) A_τ *is positive definite and belongs to* $\mathcal{S}_{\ell,u}$*;*

(Q4) conditional on the past, primary observations and auxiliary blocks are fresh copies of (X, Y)*, and the primary observation, auxiliary block, and current directions are mutually independent;*

(Q5) there is a deterministic t_0 *such that the primary trajectory lies in* $\mathcal{U}$ *almost surely for every* $t \geq t_0$*, and it satisfies* (13)*.*

The trajectory rate and localization are separate model inputs; Proposition C.42 and Remark C.43 give a risk-descent route to the rate, and the appendix gives scalar examples satisfying the conditions.

Theorem 3.5. *Under Assumption 3.4, with the direct score* (29)*, the curvature recursion* (7)*,* $\delta_t \asymp t^{-\kappa_{\mathrm{H}}}$ *and* $R_t \asymp t^r$*, if*

$$(\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha, \qquad \alpha(1+\nu) > 1, \tag{35}$$

then

$$\sqrt{n}(\bar{\theta}_n - \theta_\tau) \Longrightarrow N\big(0, \mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}\big)\,, \tag{36}$$

$$\mathrm{S}_\tau = \mathbb{E}\big[XX^{\mathsf{T}}\{\mathbf{1}(Y \leq X^{\mathsf{T}}\theta_\tau) - \tau\}^2\big]\,, \tag{37}$$

$$\widehat{\Sigma}_n \xrightarrow{p} \mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}, \tag{38}$$

$$\mathrm{S}_\tau = \tau(1-\tau)\mathbb{E}(XX^{\mathsf{T}}) \quad \textit{under} \text{ (31)}. \tag{39}$$

With independent endpoint observations, the sufficient curvature condition becomes $(2\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha$*.*

For example, with $\alpha = 3/4$ and a fixed auxiliary batch, the sufficient curvature restriction is $\kappa_{\mathrm{H}} < 1/4$ for shared endpoints and $\kappa_{\mathrm{H}} < 1/8$ for independent endpoints. At a given radius exponent, increasing r can instead offset secant variability, at the cost of more auxiliary observations. The separate condition $\alpha(1+\nu) > 1$ controls the mean-score linearization error and is unaffected by the choice of endpoint coupling. These choices concern the curvature update; the direct primary quantile score is not smoothed.

Theorem 3.6. *Suppose* $\mathrm{A} \in \mathcal{S}_{\ell,u}$*,* $\mathrm{B}_n \xrightarrow{p} \mathrm{A}$*,* $Y_t = \widehat{G}_t(\theta_{t-1})$*,* $Y_t^\star = \widehat{G}_t(\theta_\star)$*, and*

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\,\|Y_t - Y_t^\star\|^2 \longrightarrow 0, \tag{40}$$

$$\sup_{n\geq 1} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}(\|Y_t\|^2 + \|Y_t^\star\|^2) < \infty,$$

$$\frac{1}{n}\sum_{t=1}^{n} \big[Y_t^\star Y_t^{\star\mathsf{T}} - \mathbb{E}(Y_t^\star Y_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1})\big] \xrightarrow{p} 0, \tag{41}$$

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}(Y_t^\star Y_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1}) \xrightarrow{p} \mathrm{Q}. \tag{42}$$

Then $\widehat{\mathrm{Q}}_n \xrightarrow{p} \mathrm{Q}$*, and* $\widehat{\Sigma}_n \xrightarrow{p} \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$*. If, in addition, the centered conclusion of Theorem 3.2 holds, then* (11) *has asymptotic coverage* $1-\gamma$ *for every fixed contrast with positive limiting variance. A sufficient condition for* (41) *is* $\sup_t \mathbb{E}\,\|Y_t^\star\|^{2+\epsilon} < \infty$ *for some* $\epsilon > 0$*, by a standard martingale moment inequality Hall & Heyde (1980), so no fourth moment is required.*

Membership of the population curvature in the clipping set matters for sandwich inference: otherwise the Polyak-Ruppert limit may be unchanged while the learned curvature converges to its clipped version and the plug-in sandwich estimator is generally inconsistent (Corollary C.24). The appendix also verifies the theory for least squares, logistic and Huber regression (Corollaries C.10 to C.12) and analyzes smoothing bias for function-value primary updates (Theorem C.18 and Corollary C.22). Under an additional centered functional limit (Theorem C.30), self-normalization gives an alternative inference procedure that avoids estimating the asymptotic covariance (Theorem C.33).

Relation to existing methods. The transfer theorem puts the proposed recursion in the first-order class of averaged SGD, fixed or oracle preconditioning and sample Hessian stochastic Newton: every preconditioner satisfying the stabilization condition has limit $N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1})$ (Corollary C.26 and Proposition C.28). The difference is transient: as $\mathrm{B}_t \to \mathrm{A}$ the local drift $\mathrm{B}_t^{-1}\mathrm{A}$ approaches the identity, so curvature learning improves path conditioning rather than the limiting covariance, at the cost of auxiliary score queries and control of q, δ_t and R_t. Unlike sample Hessian updates, the score secant applies to Lipschitz or discontinuous scores, and unlike iterate-only random scaling it also supplies the sandwich curvature matrix; Table 4 summarizes curvature inputs, stored state and tradeoffs, and Remark C.27 relates the sandwich estimator to other online covariance estimators.

4 Numerical experiments

We compare estimation and coverage under logistic misspecification, isolate covariance estimation on a shared least squares trajectory, and examine how endpoint sharing affects quantile secant variance, each with 500 replications. RMSE is the root of the squared coefficient error averaged over coefficients and replications. Coverage is nominally 95% for the first coordinate (C) or the contrast $d^{-1/2}\sum_j \theta_j$ (R), with Monte Carlo standard error (MCSE) $\{\hat{p}(1-\hat{p})/500\}^{1/2} \approx 0.00975$ near nominal coverage. Section A gives tuning, simulation details and further mechanism diagnostics.

Misspecified logistic regression. This experiment asks whether score-based curvature learning matches Hessian-based updates under misspecification. With $d = 10$ and $n = 2 \times 10^4$, Gaussian regressors have diagonal covariance with eigenvalues decreasing geometrically from 1 to 0.01, and responses follow $\mathbb{P}(Y = 1 \mid X) = \Phi(X^\mathsf{T}\theta_{\mathrm{dgp}}/\sqrt{2})$, where θ_{dgp} repeats $(0.5, -0.25, 0.125, -0.0625)$; the fitted logistic model targets the population score root $1.14993\theta_{\mathrm{dgp}}$ rather than the generating coefficient. We compare the score secant method (proposed), sample Hessian stochastic Newton (SN), population curvature oracle Newton (ON) and SGD, all with stepsizes proportional to $t^{-0.75}$ and constants selected separately by ten-replication pilot runs of length 2000 independent of the evaluation runs, at equal numbers of primary updates. The proposed method uses three observations and five score evaluations per update (counts in Table 1); SGD covariance estimation uses population curvature as an oracle diagnostic.

In Table 1, the proposed method matches the estimation accuracy of SN and has substantially lower RMSE than ON and SGD at this horizon, with coverage reasonably close to nominal and comparable to SN across the reported contrasts; it thus avoids sample Hessians at the cost of additional observations and score evaluations.

Table 1: Misspecified logistic regression: estimation error and coverage. All 500 replications complete for every method. Observation and score counts are in thousands; SN additionally evaluates sample Hessians, and ON uses population curvature.

Method	RMSE	Cov. C	Cov. R	Obs.	Scores
proposed	0.079	0.932	0.916	60	100
SN	0.081	0.910	0.952	20	20
ON	0.245	0.000	0.462	20	20
SGD	0.252	0.282	0.418	20	20

Using the learned curvature for inference. To isolate covariance estimation, three interval procedures share one trajectory of the proposed method per replication of a least squares experiment with the regressor covariance above and $Y = X^\mathsf{T}\theta_\star + \varepsilon$, $\varepsilon \sim N(0,1)$ independent. With $d = 10$, condition number 100 and $n = 2 \times 10^4$, sandwich coordinate coverage is 0.936, versus 0.306 for nonoverlapping batch means and 0.320 for overlapping blocks, all from the same averaged estimate with RMSE 0.039; both batching rules use the fixed, unoptimized block length $\lfloor\sqrt{n}\rfloor$. Reusing the learned curvature thus improves calibration over these fixed block rules. Table 2 compares sandwich and self-normalized intervals up to $n = 5 \times 10^5$: sandwich coverage is numerically closer to nominal and mean intervals shorter, although the coverage difference at the largest sample size is not resolved by Monte Carlo precision.

Why endpoint sharing matters. Sharing observations at the two endpoints gives the upper bound $O(\delta^{-1})$, whereas independent observations give variance of order δ^{-2} under nondegeneracy (Propositions C.13 and C.14). Five quantile score experiments vary the quantile, error heterogeneity and Gaussian versus heavy-tailed observations, and pooled log variance is regressed on log radius over $\delta \in \{2^{-2}, \ldots, 2^{-7}\}$ to measure the divergence as the radius shrinks. Fitted slopes lie between -1.071 and -0.992 for shared endpoints and between -2.009 and -1.993 for independent endpoints, consistent with the theoretical bound and order respectively; the shared observation bound need not be attained in every model. Additional experiments on curvature increments (Figure 2) and equal query directional covariance (Figure 3) are reported in Sections A.2 and A.4.

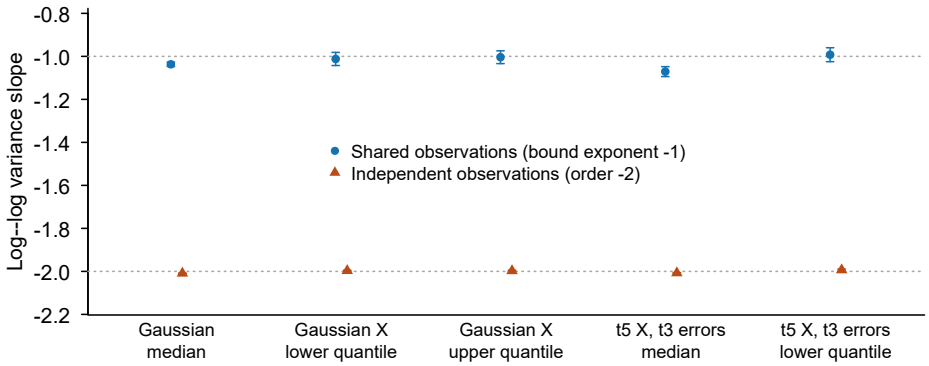


Figure 1: Quantile score secant variance slopes across five settings; shared endpoints reduce the divergence as the radius shrinks. Each point uses 500 replications, with error bars from 50 blocks of ten. Dotted lines mark slopes -1 and -2, the shared observation upper bound and the independent observation order under nondegeneracy.

Application. Section B applies the direct score quantile procedure to the UCI Online News Popularity data, fitting the median and 0.9 quantile of log popularity on 27 predictors with sandwich intervals (Table 3 and figure 4), with exact batch quantile regression on the same articles as comparator; a chronological reordering (Figure 5) changes several coefficient signs, which limits the substantive reading of the random-order estimates.

5 Concluding Remarks

We develop a stochastic Newton method that learns population curvature from paired score evaluations without sample Hessians. Under sufficient curvature stabilization the learned preconditioner is asymptotically negligible in the Polyak-Ruppert expansion, so direct unbiased scores retain the usual first-order sandwich limit; for discontinuous quantile scores, sharing observations across the secant endpoints reduces secant variability and broadens the sufficient tuning region; and when the clipping set contains the population sensitivity, the learned matrix also supports online sandwich estimation. Extensions cover function-value primary updates and self-normalized inference under additional conditions, and the experiments show these mechanisms, with performance comparable to sample Hessian stochastic Newton in the reported misspecified model but no uniform finite-sample gain established.

The theory is fixed-dimensional, relies on model-specific stability conditions and still needs auxiliary score evaluations for curvature learning. Ordered or dependent streams would need a joint score-preconditioner expansion and long-run covariance estimation, and structured secants may reduce dense matrix costs while separately controlling curvature approximation, inverse stability and matrix increments; these directions would extend the present framework to broader streaming settings.

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. When does dynamic preconditioning preserve the Polyak–Ruppert CLT? a stabilization threshold, 2026. URL https://arxiv.org/abs/2604.23498.

B. Bercu, A. Godichon, and B. Portier. An efficient stochastic Newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367, 2020.

C. Boyer and A. Godichon-Baggioni. On the asymptotic rate of convergence of stochastic Newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972, 2023.

C. Boyer and A. Godichon-Baggioni. Correction: On the asymptotic rate of convergence of stochastic Newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 87(2):705–706, 2024.

X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

X. Chen, Z. Lai, H. Li, and Y. Zhang. Online statistical inference for stochastic optimization via Kiefer–Wolfowitz methods. *Journal of the American Statistical Association*, 119(548):2972–2982, 2024.

D. Davis, D. Drusvyatskiy, and L. Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *The Annals of Statistics*, 52(4):1485–1508, 2024.

Y. Fang, J. Xu, and L. Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018. URL https://www.jmlr.org/papers/v19/17-370.html.

K. Fernandes, P. Vinagre, and P. Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Progress in Artificial Intelligence: EPIA 2015*, volume 9273 of *Lecture Notes in Computer Science*, pp. 535–546. Springer, 2015a.

K. Fernandes, P. Vinagre, P. Cortez, and P. Sernadela. Online news popularity. UCI Machine Learning Repository, 2015b. Dataset.

M. Fernandes, E. Guerre, and E. Horta. Smoothing quantile regressions. *Journal of Business & Economic Statistics*, 39(1):338–357, 2021.

A. Godichon-Baggioni and N. Werge. On adaptive stochastic optimization for streaming data: A Newton's method with $\mathcal{O}(dN)$ operations. *Journal of Machine Learning Research*, 26(59):1–49, 2025. URL https://jmlr.org/papers/v26/23-1565.html.

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

P. J. Huber. Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

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. PMLR, 2025. URL https://proceedings.mlr.press/v291/jiang25b.html.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

N. L. Kleinman, J. C. Spall, and D. Q. Naiman. Simulation-based optimization with stochastic approximation using common random numbers. *Management Science*, 45(11):1570–1578, 1999.

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

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): 7381–7389, 2022.

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(Part A):105673, 2025.

R. Leluc and F. Portier. Asymptotic analysis of conditioned stochastic gradient descent. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=U4XgzRjfF1.

T. Li, A. Kyrillidis, L. Liu, and C. Caramanis. Approximate Newton-based statistical inference using only stochastic gradients, 2018. URL https://arxiv.org/abs/1805.08920.

Yi Liu, Qirui Hu, Lei Ding, and Linglong Kong. Online local differential private quantile inference via self-normalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21698–21714. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23q.html.

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

A. Mokhtari and A. Ribeiro. RES: Regularized stochastic BFGS algorithm. *IEEE Transactions on Signal Processing*, 62(23):6089–6104, 2014.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pp. 451–459. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/hash/40008b9a5380fcacce3976bf7c08af5b-Abstract.html.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

D. Ruppert. A Newton–Raphson version of the multivariate Robbins–Monro procedure. *The Annals of Statistics*, 13(1):236–245, 1985.

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

X. Shao. A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):343–366, 2010a.

X. Shao. Corrigendum: A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):695–696, 2010b.

Y. Shen, D. Xia, and W.-X. Zhou. Online quantile regression. *Journal of Machine Learning Research*, 26(231):1–55, 2025. URL https://jmlr.org/papers/v26/24-0589.html.

J. C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45(10):1839–1853, 2000.

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

H. Wang, X. Du, and S. Na. Inference of online Newton methods with Nesterov's accelerated sketching, 2026. URL https://arxiv.org/abs/2604.23436. Accepted at ICML 2026.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. URL https://www.jstor.org/stable/1912526.

J. Zhu, L. Wang, and J. C. Spall. Efficient implementation of second-order stochastic approximation algorithms in high-dimensional problems. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):3087–3099, 2020.

W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

A Simulation details and additional results

A.1 Regression and inference specifications

All interval comparisons use nominal coverage 0.95 and $M = 500$ replications. For an estimate $\bar{\theta}_n^{(b)}$ in replication b, RMSE is $\{M^{-1}\sum_{b=1}^{M}\|\bar{\theta}_n^{(b)} - \theta_\star\|^2/d\}^{1/2}$. Coordinate coverage uses $c = (1, 0, \ldots, 0)^\mathsf{T}$, and contrast coverage uses $c = d^{-1/2}(1, \ldots, 1)^\mathsf{T}$. Reported regression configurations have 500 completed replications and no failures. Pilot and evaluation random-number streams are independent.

For the logistic comparison, $\eta_t = \eta_0 t^{-0.75}$, with $\eta_0 = 0.4$ for the proposed method and SGD, 0.2 for SN, and 0.02 for ON. Each constant minimizes pilot mean squared coefficient error over $\{0.02, 0.05, 0.1, 0.2, 0.4\}$, using ten replications of length 2000. Thus the methods are not separately optimized at the terminal evaluation horizon. Oracle curvature and the target error both use the pseudo-true logistic root; the covariance displayed for SGD also uses population curvature as an oracle diagnostic. SN estimates curvature from sample Hessians, whereas the proposed method uses paired auxiliary-score evaluations.

The proposed method starts at $\theta_0 = 0$, $\mathrm{B}_0 = \mathrm{I}_d$, uses Rademacher directions, two auxiliary observations per primary update, $\delta_t = 0.4t^{-0.15}$ and clipping interval $[0.001, 2]$. The first 50 primary updates are unpreconditioned while the curvature recursion runs. These observations remain in the lagged average (8). Each subsequent update uses the previously estimated curvature, as in (6). The two endpoints share each auxiliary observation, giving three observations and five score evaluations per primary update.

The covariance comparison uses the same Gaussian regressor distribution and coefficient vector, with linear response $Y = X^\mathsf{T}\theta_\star + \varepsilon$, $\varepsilon \sim N(0, 1)$ independent of X. The settings for the proposed method are the same except that $\eta_0 = 0.4$ and the clipping interval is $[0.01, 2]$. Within each replication, all interval methods use the same trajectory and averaged point estimate from the proposed method. Nonoverlapping batch means use blocks of length $\lfloor\sqrt{n}\rfloor$, discarding the first block and any incomplete final block. The overlapping estimator uses every complete moving block of that length and centers at the full-path mean. Both use normal critical values. The comparison concerns these fixed block implementations of batching-based covariance estimation (Chen et al., 2020; Zhu et al., 2023).

A.2 Curvature-increment rates

The median-regression diagnostic uses $d = 10$, $n = 2 \times 10^4$, Gaussian regressors of condition number 100, and independent standard Gaussian errors. The proposed method uses $\eta_0 = 0.1$, $\delta_0 = 0.4$, clipping interval $[0.001, 2]$, and 50 initial unpreconditioned updates. We record matrix increments at 32 geometrically spaced times from 10 to n, compute their root mean squared Frobenius norm across replications, and fit the log-log slope using times above the grid's 40th percentile. For quantile score curvature, we fix $q = 1$, $\alpha = 0.75$, $R_t = 2$ and compare $\kappa_\mathrm{H} = 0.20, 0.25, 0.35$. Estimated log-log increment slopes are -0.888, -0.858, and -0.818, respectively, against the upper-bound exponents -0.900, -0.875, and -0.825 in Theorem 3.3. Their ordering is consistent with increasing variability as the radius shrinks faster. All three have $s_\mathrm{H} < 1$, but only $\kappa_\mathrm{H} = 0.20$ satisfies the strict sufficient transfer condition $s_\mathrm{H} < 1 - \alpha$. The bound is not an exact finite-sample law.

A.3 Quantile endpoint coupling

For figure 1, scalar observations follow $Y = \sigma(X)\{\varepsilon - F_\varepsilon^{-1}(\tau)\}$, so the target coefficient is zero. The five settings are: Gaussian X and Gaussian errors at $\tau = 0.5$; Gaussian X and logistic errors at $\tau = 0.1, 0.9$; standardized t_5 regressors and t_3 errors at $\tau = 0.5$; and the latter distributions at $\tau = 0.1$. Heterogeneous cases use $\sigma(X) = 1 + 0.5|X|$; the two median cases use $\sigma = 1$. Each replication draws 2000 score secants at each radius $\delta = 2^{-2}, \ldots, 2^{-7}$. The shared observation scheme uses the same (X, Y) at the two endpoints, whereas the independent scheme uses separate observations. The plotted slopes regress pooled log variance on log radius; error bars use 50 blocks of ten replications.

Across these settings, shared observation slopes range from -1.071 to -0.992, while independent-endpoint slopes range from -2.009 to -1.993. The results are consistent with the shared observation

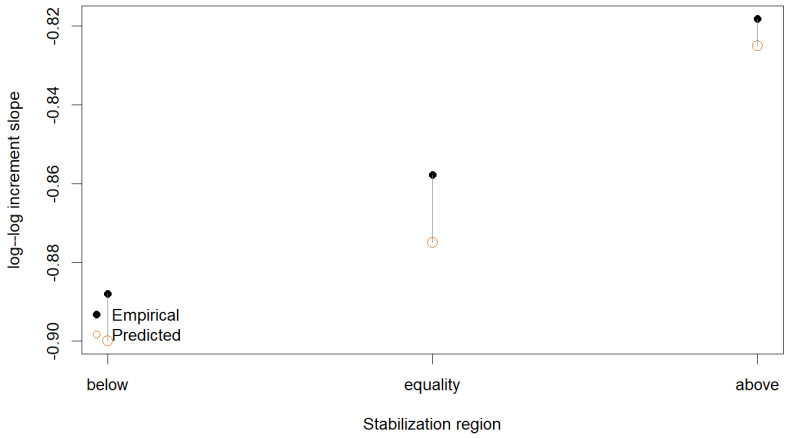


Figure 2: Estimated curvature-increment slopes and the exponents in the theorem's upper bound.

variance upper bound and the independent observation variance order under nondegeneracy in Propositions C.13 and C.14.

A.4 Directions and query budgets

The 24-cell diagnostic uses a Gaussian target score in dimension 10 with rotated covariance of condition number 20. It crosses Gaussian, Rademacher, spherical, coordinate, and orthogonal directions with $m \in \{1, 2, 5, 10\}$, together with one-sided Rademacher scores. Each of 500 replications forms the leading score $m^{-1} \sum_{j=1}^{m} U_j U_j^{\mathsf{T}} G$; the one-sided and symmetric schemes have this same leading term. Query counts are those of the corresponding function-value implementations: $2m$ per observation for symmetric scores and $m + 1$ for one-sided scores. Thus the diagnostic isolates the directional covariance in Proposition C.16, not finite-radius bias or optimization error. Relative Frobenius covariance errors range from 0.104 to 0.345; orthogonal $m = 10$ has the smallest error, 0.104. Equal-query variance indices are compared with their analytic targets.

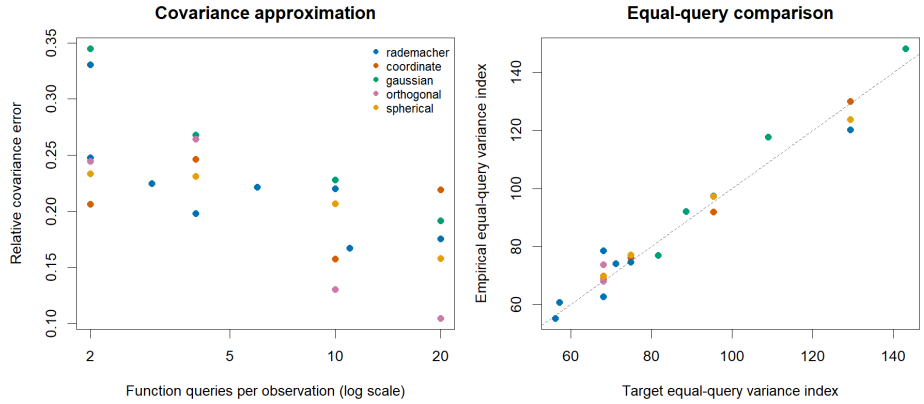


Figure 3: Directional covariance error and equal query variance.

A.5 Self-normalization and sample size

We keep the Gaussian least squares model and all settings for the proposed method in the covariance comparison fixed, and increase the number of primary updates from 2×10^4 to 100000 and 5×10^5. Each horizon uses 500 fresh evaluation streams, independent of tuning and pilot runs; matching replication indices use the same stream prefix across horizons. Neither stepsizes nor critical values are changed with the horizon. The scalar self-normalizer is computed from the lagged averages in (91), with the two-sided critical value 6.747 from Lee et al. (2022). Sandwich and self-normalized intervals share the same point estimate and differ only in calibration.

Table 2: Fixed-tuning sample-size comparison. All 500 replications complete at each horizon. Parentheses give MCSEs; coverage concerns the first coordinate. Both interval methods use the same point estimate.

n	RMSE	Sandwich coverage	Self-normalized coverage
2×10^4	0.03904 (0.00068)	0.936 (0.011)	0.858 (0.016)
100000	0.01697 (0.00027)	0.948 (0.010)	0.912 (0.013)
5×10^5	0.00720 (0.00012)	0.938 (0.011)	0.920 (0.012)

Self-normalized coverage improves with sample size but remains below nominal in Table 2. At $n = 500000$, its exact binomial 95% interval is $[0.893, 0.942]$. The sandwich-minus-self-normalized coverage difference falls from 0.078 (paired MCSE 0.015) to 0.018 (paired MCSE 0.014), so the coverage difference between the two methods is not resolved by Monte Carlo precision at the largest horizon. Mean self-normalized interval lengths are 0.03162, 0.01435, and 0.00658, compared with sandwich lengths 0.02818, 0.01244, and 0.00555. Sandwich coverage is numerically closer to nominal despite its shorter mean intervals. Self-normalization remains an asymptotic alternative that avoids covariance estimation, but increasing the sample size alone has not fully calibrated it here.

B Online News Popularity application

The application asks how selected article characteristics are associated with the conditional median and upper tail of online popularity, and whether the resulting online fit is stable to batch fitting and article order. We use the UCI Online News Popularity data Fernandes et al. (2015b;a). The sampling unit is one Mashable article. UCI's landing page and bundled names file advertise 39797 instances, whereas the official archived CSV contains 39644 rows and 61 columns with no missing values. We analyze the 39644 archived rows.

The response is $\log(1 + \text{shares})$. The model contains an intercept and 26 prespecified predictors: log-transformed word, link, image, and video counts; token and keyword summaries; six channel indicators and a weekend indicator; four LDA topic weights, with topic 4 omitted under the simplex constraint; and subjectivity, sentiment, and positive/negative word rates. Keyword-share and self-reference-share summaries are excluded because they encode popularity of other articles. Continuous variables are winsorized at the 0.5% and 99.5% quantiles and standardized using a separate 2000-article initialization block; binary indicators remain on their original scale. The resulting 27-column primary design has rank 27.

A frozen L'Ecuyer permutation allocates the remaining articles alternately to 18822 primary score and 18822 auxiliary curvature observations, with no reuse. This random-order analysis uses a working pooled exchangeability approximation. It is not a claim that publication-time observations are iid. We fit $\tau = 0.5$ and 0.9 with $\alpha = 0.75$, $\eta_0 = 0.05$, $\kappa_{\mathrm{H}} = 0.20$, $\delta_0 = 0.40$, clipping interval $[0.005, 2]$, and 200 primary updates before curvature preconditioning. Each fit processes 37644 articles, uses 56466 score queries, and stores about 4.11 MB. Final curvature eigenvalues remain within the clipping interval.

At the fitted conditional median, weekend and social-media indicators are positive, the world indicator is negative, and the content-length interval contains zero. At the 0.9 quantile, content length and the social-media, technology, and world indicators are negative, whereas global subjectivity remains positive. Exponentiating a binary coefficient gives a multiplicative contrast for the fitted conditional quantile of $1 + \text{shares}$. It is not an exact percentage change in shares, which would depend on the reference quantile. All statements are associational; neither the observational design nor the feature construction identifies causal effects.

Exact Barrodale-Roberts quantile regression on the same 18822 primary articles provides a numerical comparator. The median absolute difference between online and batch coefficients is 0.029 at $\tau = 0.5$ and 0.052 at $\tau = 0.9$; the maxima are 0.100 and 0.369. The median batch fit reports `Solution may be nonunique`. The larger upper-tail discrepancy is therefore retained rather than interpreted as numerical equivalence.

A chronological sensitivity orders articles by the UCI `timedelta` field and repeats the initialization and alternating-stream construction. It changes the sign of the median content-length coefficient and of the upper-quantile social-media, technology, world, and global-sentiment coefficients. At $\tau = 0.9$, chronological-minus-random differences are 0.537, 1.203, and 0.823 for the three channel indicators. The contrast may combine temporal composition, distribution drift, and order-dependent transient behavior; the available analysis does not identify a single cause. These changes preclude a time-stable substantive reading of the random-order coefficients.

Table 3: Selected online quantile regression coefficients under the frozen random ordering. Brackets give 95% sandwich intervals on the $\log(1 + \text{shares})$ scale.

Predictor	$\tau = 0.5$	$\tau = 0.9$
Log content words	-0.003 $[-0.074,\ 0.069]$	-0.182 $[-0.316,\ -0.048]$
Social-media channel	0.143 $[0.082,\ 0.204]$	-0.143 $[-0.248,\ -0.037]$
Technology channel	-0.061 $[-0.151,\ 0.028]$	-0.491 $[-0.583,\ -0.398]$
World channel	-0.250 $[-0.358,\ -0.141]$	-0.655 $[-0.798,\ -0.513]$
Weekend	0.283 $[0.221,\ 0.345]$	0.115 $[-0.001,\ 0.231]$
Global subjectivity	0.080 $[0.043,\ 0.117]$	0.166 $[0.064,\ 0.267]$
Global sentiment	-0.054 $[-0.095,\ -0.013]$	-0.077 $[-0.141,\ -0.014]$

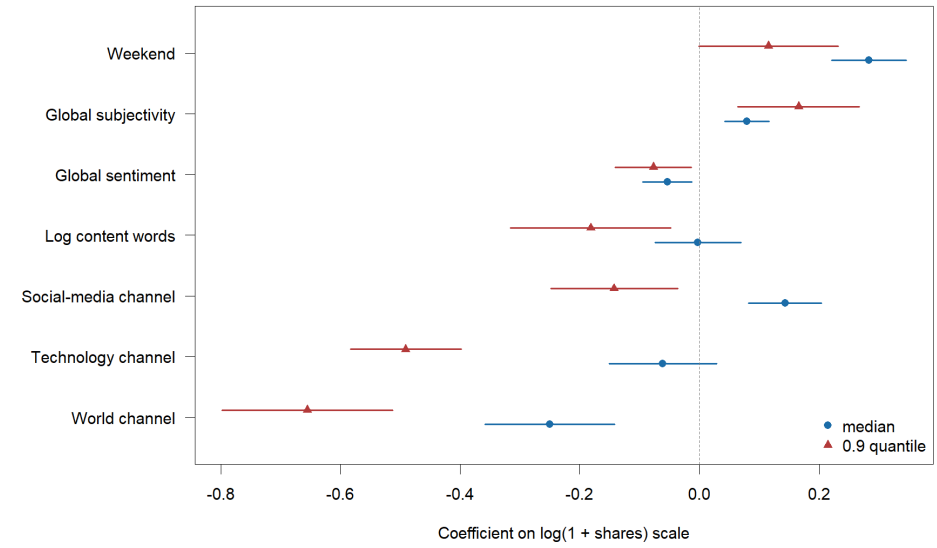


Figure 4: Online coefficient estimates and 95% sandwich intervals for selected predictors under the frozen random ordering. Continuous-variable coefficients represent one initialization-block standard deviation after winsorization.

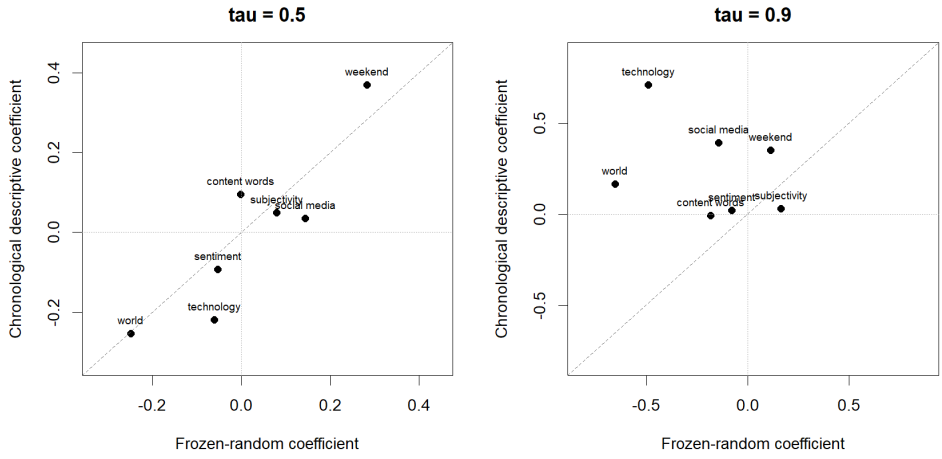


Figure 5: Chronological versus frozen random coefficient estimates. The chronological analysis is descriptive and lies outside the iid theorem.

C THEORETICAL SUPPLEMENTS AND DISCUSSION

C.1 FUNCTION-VALUE PRIMARY SCORES AND ALGORITHM

For a finite difference primary update, let $L(\theta; Z)$ be a sample loss whose paired differences have population derivative g. Draw iid isotropic directions $U_{t1}, \ldots, U_{tm}$ independently of Z_t and $\mathcal{F}_{t-1}$, unless an orthogonal-basis scheme without replacement is explicitly specified. Set

$$\widehat{G}^{(m)}_{t,h_t}(\theta) = \frac{1}{m}\sum_{k=1}^{m} \frac{L(\theta + h_t U_{tk}; Z_t) - L(\theta - h_t U_{tk}; Z_t)}{2h_t} U_{tk}. \tag{43}$$

The same primary observation is used at both endpoints and for every direction. This primary update still uses auxiliary score evaluations to learn curvature; its radius h_t controls a different error from the curvature radius δ_t.

C.2 TRANSFER CONDITIONS AND CONSEQUENCES

C.2.1 A TRANSFER THEOREM FOR APPROXIMATE NONSMOOTH SCORES

The aggregate conditions are stated in Assumption 3.1.

The six conditions are sufficient for the displayed asymptotic representation, but they are not a list of logically necessary conditions. Conditions (T1), (T3), (T4), and (T6) are standard forms of the iterate rate, linearization, local-noise, and martingale conditions used in Polyak-Ruppert theory Polyak & Juditsky (1992); Kushner & Yin (2003); Chen et al. (2020); Davis et al. (2024). Conditions (T3)–(T5) are stated as normalized sums because those sums, rather than a particular Hölder exponent, are what the proof requires. A fixed preconditioner makes (T2) automatic. For a changing preconditioner, the exponent in (T2) is the stabilization threshold of An & Huo (2026); it is sharp within the polynomial rate class used there, although additional coupling can create cancellation in a particular algorithm. A direct unbiased score makes (T5) and the target-bias term vanish. Thus Assumption 3.1 is best read as a transfer condition: a model-specific stochastic-approximation theorem may be used to verify it, and stronger pointwise assumptions should not be imposed when the aggregate terms can be controlled directly.

The identity $\mathrm{M}_t\mathrm{P}_t = \mathrm{A}^{-1}$ removes the preconditioner from the score term; its time variation remains in an Abel sum. The decomposition and polynomial stabilization threshold are due to An & Huo (2026): their Theorem 3.4 uses pathwise bounds, Proposition 3.5 treats sharpness within the stated rate class, and Proposition 5.1 also gives an L^2-moment route to a first-moment remainder bound. We give the general-β_{P} calculation used by the score secant recursion.

Proposition C.1. *Let $a_t = \eta_t^{-1}$ and define*

$$\mathcal{R}_n^{\mathrm{dyn}} := \frac{1}{\sqrt{n}} \left\{ a_1 \mathrm{M}_1 e_0 - a_n \mathrm{M}_n e_n + \sum_{t=1}^{n-1} (a_{t+1}\mathrm{M}_{t+1} - a_t \mathrm{M}_t) e_t \right\}. \tag{44}$$

Suppose $\mathbb{E}\,\|e_0\|^2 < \infty$, (13) holds, and the two bounds in (14) hold. Then $\mathbb{E}\left\|\mathcal{R}_n^{\mathrm{dyn}}\right\| \longrightarrow 0$, hence $\mathcal{R}_n^{\mathrm{dyn}} = o_p(1)$. The exponent $(1+\alpha)/2$ cannot be lowered using only the corresponding boundedness, iterate rate, and polynomial stabilization hypotheses: the construction of An and Huo that attains the rate (An & Huo, 2026, Proposition 3.5) already lies in this class. This does not exclude cancellation created by additional coupling in a particular adaptive recursion.

The main result is stated in Theorem 3.2.

Remark C.2 (Interpretation)**.** *The preconditioner affects the trajectory and hence finite-sample behavior, but it does not appear in the influence function once the dynamic Abel term is negligible. The theorem extends the established smooth unbiased dynamic-preconditioning result to time-varying approximate scores by separating three additional terms: target bias, local bias variation, and local nonsmooth score replacement.*

The aggregate formulation avoids imposing a stronger pointwise assumption than the proof uses. The next proposition gives a familiar sufficient route.

Algorithm 1: Online secant Newton estimation and inference

Require: Initial values $\theta_0 \in \mathbb{R}^d$ and $\mathrm{B}_0 \in \mathcal{S}_{\ell,u}$; stepsizes η_t; secant radii δ_t; auxiliary batch sizes R_t; and, for the finite difference primary score in (43), radii h_t and direction count m. Set $\bar{\theta}_0 = 0$ and $\widehat{\mathrm{Q}}_0 = 0$.
for $t = 1, 2, \ldots$ **do**

1. Observe Z_t and compute $\widehat{G}_t(\theta_{t-1})$ using either the direct score $G(\theta_{t-1}; Z_t)$ or the finite difference score in (43).
2. Solve $\mathrm{B}_{t-1} s_t = \widehat{G}_t(\theta_{t-1})$ for $s_t \in \mathbb{R}^d$, and update the parameter by $\theta_t = \theta_{t-1} - \eta_t s_t$.
3. Update the averaged estimator: $\bar{\theta}_t = \bar{\theta}_{t-1} + t^{-1}(\theta_{t-1} - \bar{\theta}_{t-1})$.
4. Draw a direction V_t and auxiliary observations $W_{t1}, \ldots, W_{tR_t}$ according to the sampling scheme specified above. In the direct score setting, the current observation may be reused by setting $W_{t1} = Z_t$ under Proposition C.8.
5. Evaluate $\Psi(\theta_{t-1} + \delta_t V_t; W_{tj})$ and $\Psi(\theta_{t-1} - \delta_t V_t; W_{tj})$ for $j = 1, \ldots, R_t$, using the same observation at both endpoints. Construct $\widehat{\mathrm{H}}_t$ from (4)–(5), and update B_t by (7). The updated matrix is first used in the parameter update at iteration $t + 1$.
6. Update the score second moment estimate: $\widehat{\mathrm{Q}}_t = (1 - t^{-1})\widehat{\mathrm{Q}}_{t-1} + t^{-1}\widehat{G}_t(\theta_{t-1})\widehat{G}_t(\theta_{t-1})^{\mathsf{T}}$.

Output: After n iterations, report $\bar{\theta}_n$. Compute the sandwich estimate $\widehat{\Sigma}_n = \mathrm{B}_n^{-1}\widehat{\mathrm{Q}}_n\mathrm{B}_n^{-1}$ using linear solves with B_n, and use $\widehat{\Sigma}_n/n$ for covariance estimation. Under the conditions of Theorems 3.2 and 3.6 and the centered target-bias condition, report the confidence interval in (11).

Figure 6: Primary and curvature update order and reporting.

Proposition C.3. *Suppose $\nu \in (0, 1]$, $\chi, \xi \in (0, 2]$, and let $\{\rho_t\}$ and $\{L_t\}$ be nonnegative deterministic sequences. Suppose either that the following inequalities hold globally, or that they hold on one common neighborhood of $\theta_\star$ and there is a deterministic t_0 such that, almost surely, θ_{t-1} belongs to that neighborhood for every $t \geq t_0$. In the local alternative, assume also that the finitely many prefix terms appearing in* (15)–(17) *are integrable. Assume*

$$\|r(e)\| \leq C \|e\|^{1+\nu}, \tag{45}$$

$$\mathbb{E}\{\|\zeta_t(\theta) - \zeta_t(\theta_\star)\|^2 \mid \mathcal{F}_{t-1}\} \leq C \|\theta - \theta_\star\|^\chi + \rho_t, \qquad \frac{1}{n}\sum_{t=1}^{n} \rho_t \to 0, \tag{46}$$

$$\|b_t(\theta) - b_t(\theta_\star)\| \leq L_t \|\theta - \theta_\star\|^\xi, \qquad \frac{1}{\sqrt{n}}\sum_{t=1}^{n} L_t \eta_{t-1}^{\xi/2} \to 0. \tag{47}$$

If (13) *holds and*

$$\alpha(1 + \nu) > 1, \tag{48}$$

then (15)*,* (16)*, and* (17) *hold. Condition* (48) *is a convenient sufficient condition for the mean remainder, not a necessary part of the transfer theorem; an unbiased direct score makes* (47) *automatic.*

This proposition is a verification result, not an additional assumption of the transfer theorem, and none of its three pointwise inequalities is individually necessary. The expansion $r(e) = O(\|e\|^{1+\nu})$ is the usual local remainder in averaged stochastic approximation; $\nu = 1$ follows from a locally Lipschitz Jacobian, while smaller ν covers weaker differentiability. The restriction $\alpha(1 + \nu) > 1$ is only the consequence of combining that bound with the generic mean-square rate $\mathbb{E}\|e_t\|^2 = O(t^{-\alpha})$. It is automatic for $\nu = 1$ and every $\alpha > 1/2$, and it can be replaced by direct verification of (T3). The quadratic-mean score condition is a standard way to handle nonsmooth stochastic scores Polyak & Juditsky (1992); Davis et al. (2024); the local bias condition is absent for an unbiased score. Similar local expansions and moment controls appear in online SGD inference Chen et al. (2020); Lee et al. (2022). The proposition is retained because it reduces the aggregate conditions to familiar model calculations without making higher-order smoothness part of the main theorem.

C.2.2 MOMENT BOUNDS FOR THE SCORE SECANT

Assumption C.4. *There are bounded neighborhoods $\mathcal{U} \subset \mathcal{U}_+$ of $\theta_\star$, with $\overline{\mathcal{U}_+}$ compact, a radius $\bar{\delta} > 0$, and a symmetric positive definite matrix $\mathrm{A} = Dg(\theta_\star)$ such that:*

(S1) The function g is continuously differentiable on an open set containing $\overline{\mathcal{U}_+}$, and $\mathrm{H}(\theta) = Dg(\theta)$ is symmetric. Define its modulus of continuity on $\overline{\mathcal{U}_+}$ by

$$\omega_{\mathrm{H}}(a) = \sup\left\{\|\mathrm{H}(\theta_1) - \mathrm{H}(\theta_2)\|_{\mathrm{F}} : \theta_1, \theta_2 \in \overline{\mathcal{U}_+},\ \|\theta_1 - \theta_2\| \le a\right\}.$$

Then $\omega_{\mathrm{H}}(a) \to 0$ as $a \downarrow 0$. This requires no derivative of the population Jacobian and no sample Hessian.

(S2) Conditional on $\mathcal{F}_{t-1}$, the observations $W_{t1}, \dots, W_{tR_t}$ are independent copies of W and are independent of V_t. The direction is independent of the past, satisfies $\mathbb{E}(V_t V_t^{\mathsf{T}}) = \mathrm{I}_d$, and $\|V_t\| \le \bar{v}$ almost surely. Moreover, $\theta + s\delta v \in \mathcal{U}_+$ for every $\theta \in \mathcal{U}$, $s \in [-1, 1]$, $0 < \delta \le \bar{\delta}$, and v in the support of V_t. Scaled spherical, Rademacher, and coordinate directions satisfy these conditions.

(S3) For some $q \ge 0$, every $\theta \in \mathcal{U}$, and $0 < \delta \le \bar{\delta}$,

$$\mathbb{E}\Big[\|\Psi(\theta + \delta V; W) - \Psi(\theta - \delta V; W)\|^2 \|V\|^2\Big] \le C\delta^{2-q}. \tag{49}$$

Condition (S1) is a first-order population condition: continuity of Dg on a compact perturbation tube is enough for consistency of the secant mean, and no derivative of Dg is used. Symmetry is appropriate when g is the gradient of a locally convex risk; a nonsymmetric estimating equation requires a different regularization of its Jacobian. Bounded isotropic directions and fresh auxiliary observations in (S2) are convenient sufficient choices, not necessary features of the method. Unbounded directions can be used with an integrable domination condition. Lagged reuse is admissible after its conditional moments are checked, and Proposition C.8 treats one current direct score observation. Condition (S3) is not a generic smoothness assumption but a model-specific second moment bound. Perturbation-based Hessian estimation routinely imposes analogous moment controls Spall (2000); Zhu et al. (2020); the value of the present formulation is that the exponent q is calculated rather than hidden in a uniform variance assumption.

Proposition C.5. *Under Assumption C.4, on the event $\{\theta_{t-1} \in \mathcal{U}\}$ the curvature observation* (5) *satisfies, conditionally on $\mathcal{F}_{t-1}$,*

$$\left\|\mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1}) - \mathrm{A}\right\|_{\mathrm{F}} \le C\left\{\omega_{\mathrm{H}}(\|e_{t-1}\|) + \omega_{\mathrm{H}}(\bar{v}\delta_t)\right\}, \tag{50}$$

$$\mathbb{E}\left[\left\|\widehat{\mathrm{H}}_t - \mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1})\right\|_{\mathrm{F}}^2 \mid \mathcal{F}_{t-1}\right] \le C\left(1 + \frac{1}{R_t \delta_t^q}\right). \tag{51}$$

Proposition C.6. *Suppose, on the perturbation tube,*

$$\|\Psi(\theta_1; W) - \Psi(\theta_2; W)\| \le L(W)\|\theta_1 - \theta_2\|, \qquad \mathbb{E}L(W)^2 < \infty. \tag{52}$$

If V is bounded, then (49) *holds with $q = 0$. Thus a score may be nonsmooth and still have $q = 0$; pathwise Lipschitz continuity is enough.*

Remark C.7 (Interpretation of q)**.** *For a Lipschitz score, the numerator of the secant is of order δ, so $q = 0$ is admissible. For a jump score evaluated with the same observation at both endpoints, the numerator is nonzero only when the perturbation crosses a discontinuity; an $O(\delta)$ crossing probability makes $q = 1$ admissible. Independent noisy endpoint evaluations generally retain a nonvanishing difference variance as $\delta \downarrow 0$, which gives the exact exponent $q = 2$ under nondegeneracy.*

The main result is stated in Theorem 3.3.

Proposition C.8. *Suppose the primary oracle is the direct score $G(\theta; Z_t)$. Conditional on $\mathcal{F}_{t-1}$, let $Z_t, W_{t2}, \dots, W_{tR_t}, V_t$ be mutually independent, let $Z_t, W_{t2}, \dots, W_{tR_t}$ have the auxiliary observation law, and let V_t satisfy the direction and perturbation-tube requirements in (S2). Set $W_{t1} = Z_t$ and use the same Z_t at the two endpoints of its secant. Compute the primary update with $\mathrm{P}_t = \mathrm{B}_{t-1}^{-1}$, center every secant at θ_{t-1}, and use the resulting B_t first through $\mathrm{P}_{t+1} = \mathrm{B}_t^{-1}$.*

Then P_t is predictable for the primary update, and the conditional mean and variance conclusions in Proposition C.5 are unchanged. Consequently, Theorems 3.2 and 3.3 remain valid when their other conditions hold. The same reuse does not alter Theorem 3.6 when its stated outer-product conditions hold.

Remark C.9 (The boundary case)**.** *Condition* (28) *is sufficient. At equality, the general Abel bound is of order one at the root-n scale. The example of An and Huo that attains the rate An & Huo (2026) shows that the threshold cannot be improved from the stated rate assumptions alone. A more favorable boundary limit would have to use additional structure of a particular curvature recursion.*

C.2.3 Applications to common regression problems

The following results verify the transfer conditions for familiar fixed-dimensional M-estimation scores; see, for example, van der Vaart (1998). Least squares admits a primitive rate proof, whereas the logistic and Huber results retain the mean-square iterate rate as a separate stability input. In all three applications, the initialization has a finite second moment about the stated target, primary observations and auxiliary batches are independent across time and between streams, $\mathrm{P}_t = \mathrm{B}_{t-1}^{-1}$ is formed by (7), $R_t \equiv R \geq 1$, and $\delta_t = \delta_0 t^{-\kappa_{\mathrm{H}}}$ for any $\kappa_{\mathrm{H}} > 0$. The clipping bounds are chosen so that the population curvature belongs to $\mathcal{S}_{\ell,u}$.

Corollary C.10. *Let*

$$Y = X^{\mathsf{T}}\theta_{\mathrm{LS}} + \varepsilon, \qquad \mathbb{E}(\varepsilon \mid X) = 0, \qquad G_{\mathrm{LS}}(\theta; X, Y) = X(X^{\mathsf{T}}\theta - Y).$$

Assume

$$\mathrm{A}_{\mathrm{LS}} = \mathbb{E}(XX^{\mathsf{T}}) \in \mathcal{S}_{\ell,u}, \qquad \mathbb{E}\,\|X\|^4 < \infty, \qquad \mathbb{E}\,\|\varepsilon X\|^{2+\epsilon_0} < \infty \tag{53}$$

for some $\epsilon_0 > 0$. For sufficiently small η_0,

$$\sqrt{n}(\bar{\theta}_n - \theta_{\mathrm{LS}}) \Longrightarrow N(0, \Gamma_{\mathrm{LS}}), \tag{54}$$

$$\Gamma_{\mathrm{LS}} = \mathrm{A}_{\mathrm{LS}}^{-1}\mathrm{Q}_{\mathrm{LS}}\mathrm{A}_{\mathrm{LS}}^{-1}, \qquad \mathrm{Q}_{\mathrm{LS}} = \mathbb{E}(\varepsilon^2 XX^{\mathsf{T}}), \tag{55}$$

$$\widehat{\Sigma}_n \xrightarrow{p} \Gamma_{\mathrm{LS}}. \tag{56}$$

The secant is unbiased for A_{LS} for every $\delta_t > 0$ and has $q = 0$. Under conditional homoskedasticity, $\mathbb{E}(\varepsilon^2 \mid X) = \sigma^2$, the covariance reduces to $\Gamma_{\mathrm{LS}} = \sigma^2 \mathrm{A}_{\mathrm{LS}}^{-1}$.

Corollary C.11. *Let $Y \in \{0, 1\}$,*

$$p_\theta(x) = \{1 + \exp(-x^{\mathsf{T}}\theta)\}^{-1}, \qquad G_{\log}(\theta; X, Y) = X\{p_\theta(X) - Y\},$$

and let $\theta_{\log}$ solve $\mathbb{E}G_{\log}(\theta; X, Y) = 0$. Correct logistic specification is not required. Suppose $\mathbb{E}\,\|X\|^4 < \infty$ and there is a convex neighborhood $\mathcal{U}$ on which

$$\mathrm{H}_{\log}(\theta) = \mathbb{E}\{p_\theta(X)(1 - p_\theta(X))XX^{\mathsf{T}}\} \succeq \mu \mathrm{I}_d, \qquad \mathrm{A}_{\log} = \mathrm{H}_{\log}(\theta_{\log}) \in \mathcal{S}_{\ell,u} \tag{57}$$

for some $\mu > 0$. Assume that the unmodified primary trajectory satisfies (13)*. This rate is a separate stability input; local positive curvature does not by itself establish it. Global score bounds in the proof avoid imposing deterministic-time confinement. Section C.5.9 gives an attainable scalar example. Then*

$$\sqrt{n}(\bar{\theta}_n - \theta_{\log}) \Longrightarrow N(0, \Gamma_{\log}), \tag{58}$$

$$\Gamma_{\log} = \mathrm{A}_{\log}^{-1}\mathrm{Q}_{\log}\mathrm{A}_{\log}^{-1}, \qquad \mathrm{Q}_{\log} = \mathbb{E}\big[\{Y - p_{\theta_{\log}}(X)\}^2 XX^{\mathsf{T}}\big], \tag{59}$$

$$\widehat{\Sigma}_n \xrightarrow{p} \Gamma_{\log}. \tag{60}$$

The score is pathwise Lipschitz and its secant has $q = 0$. Under the correctly specified Bernoulli model, $\mathrm{Q}_{\log} = \mathrm{A}_{\log}$ and $\Gamma_{\log} = \mathrm{A}_{\log}^{-1}$; see McCullagh & Nelder (1989) for the standard likelihood identities.

Corollary C.12. *Let $\psi_c(u) = \max\{-c, \min(c, u)\}$ and*

$$G_c(\theta; X, Y) = X\psi_c(X^{\mathsf{T}}\theta - Y), \qquad \mathbb{E}G_c(\theta_c; X, Y) = 0.$$

Suppose $\mathbb{E}\,\|X\|^4 < \infty$, and let $\mathcal{U}$ be a convex neighborhood of θ_c. Assume the conditional density exists and is bounded on the two threshold tubes $X^{\mathsf{T}}\theta \pm c$, $\theta \in \mathcal{U}$, by an envelope $\bar{f}_c(X)$ satisfying $\mathbb{E}\{\bar{f}_c(X)\,\|X\|^3\} < \infty$. Assume also that, on $\mathcal{U}$,

$$\mathrm{H}_c(\theta) = \mathbb{E}\big[XX^{\mathsf{T}}\mathbb{P}\{|Y - X^{\mathsf{T}}\theta| < c \mid X\}\big] \succeq \mu \mathrm{I}_d, \qquad \mathrm{A}_c = \mathrm{H}_c(\theta_c) \in \mathcal{S}_{\ell,u}. \tag{61}$$

Suppose the unmodified primary trajectory satisfies (13)*. This rate is a separate stability input. The local density conditions control the population expansion near the target, while global bounded-score and Lipschitz bounds control excursions outside that neighborhood. Then*

$$\sqrt{n}(\bar{\theta}_n - \theta_c) \Longrightarrow N(0, \Gamma_c), \tag{62}$$

$$\Gamma_c = \mathrm{A}_c^{-1}\mathrm{Q}_c\mathrm{A}_c^{-1}, \qquad \mathrm{Q}_c = \mathbb{E}\{\psi_c(X^\mathsf{T}\theta_c - Y)^2 XX^\mathsf{T}\}, \tag{63}$$

$$\widehat{\Sigma}_n \xrightarrow{p} \Gamma_c. \tag{64}$$

The Huber score is nonsmooth but one-Lipschitz in the residual, so its secant has $q = 0$. The density condition is used for differentiability and local positive curvature of the population risk, as in classical robust regression Huber (1973); no derivative of the density is required.

The population sensitivity matrices, score covariances, and secant exponents are summarized in Table 5.

C.2.4 Quantile regression and endpoint coupling

The quantile setup and model conditions are stated in Assumption 3.4.

The density assumption is a first-order local condition: it requires no derivative of $f_{Y|X}$. The three moment conditions control, respectively, the target score, the probability of crossing the indicator discontinuity, and the population-score remainder. When $\nu = 1$, the later restriction $\alpha(1+\nu) > 1$ is automatic for every $\alpha > 1/2$.

Use $\Psi = G_\tau$ in (4)–(5), with the same (X, Y) at the positive and negative endpoints.

Proposition C.13. *Under Assumption 3.4, once the center and perturbations lie in the specified tube,*

$$\left\|\mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1}) - \mathrm{A}_\tau\right\|_{\mathrm{F}} \le C\{\|\theta_{t-1} - \theta_\tau\|^\nu + \delta_t^\nu\}, \tag{65}$$

$$\mathbb{E}\left[\left\|\widehat{\mathrm{H}}_t - \mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1})\right\|_{\mathrm{F}}^2 \mid \mathcal{F}_{t-1}\right] \le C\left(1 + \frac{1}{R_t\delta_t}\right). \tag{66}$$

Hence $q = 1$ is admissible in (49)*.*

Proposition C.14. *Assume (Q1)–(Q2) of Assumption 3.4 and fix $\theta \in \mathcal{U}$. Let Z^+ and Z^- be independent copies of Z, independent of a bounded isotropic direction V, and define*

$$\mathrm{K}_\delta^{\mathrm{ind}} = \mathrm{sym}\left(\frac{G_\tau(\theta + \delta V; Z^+) - G_\tau(\theta - \delta V; Z^-)}{2\delta} V^\mathsf{T}\right). \tag{67}$$

Uniformly for θ in the parameter tube,

$$\mathbb{E}\left\|\mathrm{K}_\delta^{\mathrm{ind}}\right\|_{\mathrm{F}}^2 \le C\delta^{-2}. \tag{68}$$

If $\mathrm{tr}\{\mathrm{Var}(G_\tau(\theta; Z))\} > 0$, then

$$\delta^2\mathbb{E}\left\|\mathrm{K}_\delta^{\mathrm{ind}} - \mathbb{E}\mathrm{K}_\delta^{\mathrm{ind}}\right\|_{\mathrm{F}}^2 \longrightarrow \frac{d+1}{4}\,\mathrm{tr}\{\mathrm{Var}(G_\tau(\theta; Z))\} > 0. \tag{69}$$

Thus the sampling component of a batch average is of order $(R_t\delta_t^2)^{-1}$ and $q = 2$.

The main result is stated in Theorem 3.5.

Remark C.15 (Relation to online quantile regression)**.** *Direct stochastic-subgradient methods for online quantile regression are already available, including increasing-dimensional results Shen et al. (2025). The additional result here concerns estimation of curvature from coupled score evaluations and the resulting restriction on δ_t and R_t.*

C.2.5 Zeroth-order primary scores and smoothing bias

Let $G_\star = G(\theta_\star; Z)$ be a mean-zero direct score with $\mathrm{S} = \mathbb{E}(G_\star G_\star^\mathsf{T})$. For one isotropic direction, define

$$\mathrm{Q}_1 = \mathbb{E}_U(UU^\mathsf{T}\mathrm{S}UU^\mathsf{T}). \tag{70}$$

Proposition C.16 (Known random-direction covariance; Chen et al. Chen et al. (2024))**.** *Suppose the finite difference score in* (43) *converges in L^2 at the target to $m^{-1}\sum_{k=1}^{m} U_k U_k^{\mathsf{T}} G_\star$. Let the directions be iid, independent of $G_\star$ and the past, with $\mathbb{E}(U_k U_k^{\mathsf{T}}) = \mathrm{I}_d$ and finite Q_1 in* (70)*. Then its limiting covariance is $\mathrm{Q}_m = \frac{1}{m}\mathrm{Q}_1 + \frac{m-1}{m}\mathrm{S}$. For $d \geq 2$ and $1 \leq m \leq d$, scaled orthogonal-basis directions sampled without replacement have covariance*

$$\mathrm{Q}_m^{\mathrm{WOR}} = \frac{d-m}{m(d-1)}\mathrm{Q}_1 + \frac{d(m-1)}{m(d-1)}\mathrm{S}.$$

For the same marginal orthogonal-basis distribution, this matrix is no larger than its iid counterpart Q_m in Loewner order and equals S when $m = d$. These formulas are the multi-query directional covariance calculations of Chen et al. (2024). Their one-sided scheme has different query accounting from the $2m$ symmetric function queries used here.

For the deterministic smoothing bias, suppose

$$b_t(\theta_\star) = h_t^p B + \rho_t^b, \qquad h_t = h_0 t^{-\kappa}, \qquad \mathbb{E}\left\|\rho_t^b\right\| = o(h_t^p), \tag{71}$$

for $p > 0$ and $B \in \mathbb{R}^d$. The expansion is stronger than a mere order bound and is needed only to identify the critical mean shift.

Proposition C.17. *For $0 < h \leq h_0$, write the population smoothing bias as $b_h(\theta) = \mathbb{E}\{\widehat{G}_h^{(m)}(\theta)\} - g(\theta)$. Suppose that, on a neighborhood $\mathcal{U}$ of $\theta_\star$,*

$$b_h(\theta) = h^p B(\theta) + r_h(\theta), \qquad \|B(\theta) - B(\theta_\star)\| \leq C\,\|\theta - \theta_\star\|^{\xi}, \qquad \sup_{\theta\in\mathcal{U}} \|r_h(\theta)\| \leq h^p \omega(h), \tag{72}$$

where $p > 0$, $\xi \in (0, 2]$, and $\omega(h) \to 0$. If the trajectory lies in $\mathcal{U}$ almost surely after a deterministic time t_0, satisfies (13)*, and*

$$\sum_{t<t_0} \mathbb{E}\,\|b_{h_t}(\theta_{t-1}) - b_{h_t}(\theta_\star)\| < \infty,$$

then

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \mathbb{E}\,\|b_{h_t}(\theta_{t-1}) - b_{h_t}(\theta_\star)\| \leq \frac{C}{\sqrt{n}}\sum_{t=1}^{n} h_t^p \eta_{t-1}^{\xi/2} + \frac{2}{\sqrt{n}}\sum_{t=1}^{n} h_t^p \omega(h_t) + o(1). \tag{73}$$

In particular, for $h_t = h_0 t^{-\kappa}$, condition (17) *holds throughout the centered and critical regimes $p\kappa \geq 1/2$. More generally, it holds whenever both sums on the right side of* (73) *tend to zero.*

Theorem C.18. *Assume the transfer conditions hold with target covariance Q_m, the local bias condition* (17) *is valid, and* (71) *holds. Then:*

(i) if $p\kappa > 1/2$,

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N(0, \mathrm{A}^{-1}\mathrm{Q}_m\mathrm{A}^{-\mathsf{T}});$$

(ii) if $p\kappa = 1/2$,

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N(-2h_0^p\mathrm{A}^{-1}B, \mathrm{A}^{-1}\mathrm{Q}_m\mathrm{A}^{-\mathsf{T}});$$

(iii) if $\alpha/2 \leq p\kappa < 1/2$ and $B \neq 0$, the root-n sequence centered at $\theta_\star$ is not tight; its deterministic centering has order $n^{1/2-p\kappa}$.

Under the other aggregate transfer conditions, if $p\kappa < \alpha/2$ and $B \neq 0$, (T1) is incompatible with the target-bias expansion; no conclusion is asserted in that range.

Remark C.19 (Interpretation of the bias regimes)**.** *The smoothing rate changes the centering but not the stochastic covariance. In particular, the sandwich estimator remains a covariance estimator at the critical rate, but an uncorrected Wald interval is shifted. Undersmoothing avoids estimation of B; critical-rate inference requires either an explicit B or a valid bias correction.*

Proposition C.20. *Suppose, for a fixed θ_{ref}, the anchored population risk $R(\theta) = \mathbb{E}\{L(\theta; Z) - L(\theta_{\mathrm{ref}}; Z)\}$ is finite and differentiable near $\theta_\star$, with $g = \nabla R$, and expectation may be interchanged with the paired symmetric difference. A finite unanchored expected loss is a sufficient special case;*

the anchor cancels from every paired difference. Assume that g is differentiable at $\theta_\star$, with derivative A*, in the form*

$$\|g(\theta_\star + v) - \mathrm{A}v\| \le \|v\|\ \omega_g(\|v\|), \qquad \omega_g(a) \longrightarrow 0 \quad (a \downarrow 0).$$

Let U have bounded support, $\|U\| \le \bar{u}$. Then

$$\left\| \mathbb{E}\widehat{G}^{(m)}_{t,h}(\theta_\star) - g(\theta_\star) \right\| \le Ch\,\omega_g(\bar{u}h) = o(h). \tag{74}$$

Hence a centered root-n limit follows whenever

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} h_t \omega_g(\bar{u}h_t) \longrightarrow 0, \tag{75}$$

provided the remaining transfer conditions hold. If $\omega_g(a) \le Ca^\nu$ for some $\nu \in (0, 1]$, then

$$\left\| \mathbb{E}\widehat{G}^{(m)}_{t,h_t}(\theta_\star) - g(\theta_\star) \right\| = O(h_t^{1+\nu}), \tag{76}$$

and $\kappa(1 + \nu) > 1/2$ is sufficient when $h_t \asymp t^{-\kappa}$. For unbounded directions, the same conclusions require a corresponding integrable domination along the random line segment. Identifying a critical mean shift still requires a sharper expansion.

The differentiable-density conditions used next are stronger and are needed only to identify the critical bias constant. Comparable density expansions are standard in smoothed quantile regression theory Fernandes et al. (2021).

Proposition C.21. *Under the risk-score identity, expectation-interchange condition, and bounded-direction condition of Proposition C.20, suppose, for a symmetric bilinear map $\mathcal{T} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$,*

$$g(\theta_\star + v) = \mathrm{A}v + \frac{1}{2}\mathcal{T}[v, v] + r_2(v), \qquad \|r_2(v)\| = o(\|v\|^2), \tag{77}$$

uniformly over the bounded direction support. Then the population bias of the symmetric zeroth-order score satisfies

$$b_h(\theta_\star) = h^2 B + o(h^2), \qquad B = \frac{1}{6}\mathbb{E}_U\{UU^\mathsf{T}\mathcal{T}[U, U]\}. \tag{78}$$

For quantile regression, suppose additionally that, with $q_\star = X^\mathsf{T}\theta_\tau$,

$$\left| f_{Y|X}(q_\star + s \mid X) - f_{Y|X}(q_\star \mid X) - \dot{f}_\star(X)s \right| \le L_2(X)\,|s|^{1+\gamma} \tag{79}$$

for every covariate-scaled shift $s = X^\mathsf{T}v$ from the permitted parameter perturbations, with some $\gamma > 0$ and an integrable envelope satisfying

$$\mathbb{E}\left[(\left|\dot{f}_\star(X)\right| + L_2(X))(1 + \|X\|^{3+\gamma}) \right] < \infty. \tag{80}$$

Then (77) *holds with*

$$\mathcal{T}_\tau[v, v] = \mathbb{E}\{\dot{f}_\star(X)X(X^\mathsf{T}v)^2\}, \tag{81}$$

and the critical-bias vector is

$$B_\tau = \frac{1}{6}\mathbb{E}\{\dot{f}_\star(X)UU^\mathsf{T}X(X^\mathsf{T}U)^2\}. \tag{82}$$

If, in addition, (79) *holds uniformly with $q_\star$ replaced by $q_\theta = X^\mathsf{T}\theta$ for θ in a neighborhood, with derivative $\dot{f}_\theta(X)$, and*

$$\left| \dot{f}_{\theta_1}(X) - \dot{f}_{\theta_2}(X) \right| \le L_{\dot{f}}(X)\left| X^\mathsf{T}(\theta_1 - \theta_2) \right|^\xi, \qquad \mathbb{E}\{L_{\dot{f}}(X)(1 + \|X\|^{3+\xi})\} < \infty, \tag{83}$$

then (72) *holds with $p = 2$,*

$$B_\tau(\theta) = \frac{1}{6}\mathbb{E}\{\dot{f}_\theta(X)UU^\mathsf{T}X(X^\mathsf{T}U)^2\}, \tag{84}$$

$\|B_\tau(\theta) - B_\tau(\theta_\tau)\| \le C\,\|\theta - \theta_\tau\|^\xi$*, and* $\omega(h) = O(h^\gamma)$.

Corollary C.22. *Under Assumption 3.4, use the check loss in* (43)*, a fixed number of iid bounded isotropic primary directions, independent of the primary observation and past, and the same score secant preconditioner as in Theorem 3.5. Condition (Q5), including localization and the mean-square rate, is imposed on this zeroth-order trajectory. The open convex tube in (Q1) also contains all primary segments $\theta + sh_tU_{tk}$, $s \in [-1, 1]$, used by this trajectory. Suppose the target-bias expansion holds for an order $p \geq 1 + \nu$, with coefficient B possibly zero. Assume* (17)*, or verify it through Proposition C.17; also require both $(\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha$ and $\alpha(1 + \nu) > 1$ in* (35)*. Then the covariance in* (36) *is replaced by $\mathrm{A}_\tau^{-1}\mathrm{Q}_m\mathrm{A}_\tau^{-1}$, with Q_m from Proposition C.16, and the three regimes are determined by $p\kappa$ relative to $1/2$, subject to the compatibility qualification in Theorem C.18. Under* (79)–(80)*, one may take $p = 2$ and $B = B_\tau$ from* (82)*. The uniform extension* (83) *verifies the local-bias condition automatically when $2\kappa \geq 1/2$. In the centered and critical regimes, $\widehat{\Sigma}_n \xrightarrow{p} \mathrm{A}_\tau^{-1}\mathrm{Q}_m\mathrm{A}_\tau^{-1}$. At the critical rate this estimates the stochastic covariance, but the Wald interval must also correct the nonzero mean shift.*

Remark C.23 (Two independent tuning restrictions)**.** *The curvature radius δ_t controls stabilization of B_t through $(q\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha$. The primary smoothing radius h_t controls centering through $p\kappa > 1/2$. They solve different problems and should not be set equal by default.*

C.2.6 Online covariance estimation and special cases

The main result is stated in Theorem 3.6.

Corollary C.24. *Let A be symmetric positive definite and let $\mathrm{A}_{\ell,u} = \Pi_{\ell,u}(\mathrm{A})$. Suppose (T1) and (T3)–(T6) hold, the target bias in* (20) *is zero, and the assumptions of Theorem 3.3 hold with $s_{\mathrm{H}} < 1 - \alpha$. Also suppose all conditions of Theorem 3.6 on Y_t, $Y_t^\star$, and Q hold, without its premises $\mathrm{A} \in \mathcal{S}_{\ell,u}$ and $\mathrm{B}_n \xrightarrow{p} \mathrm{A}$. Then*

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\zeta_t^\star + o_p(1),$$
$$\widehat{\Sigma}_n \xrightarrow{p} \mathrm{A}_{\ell,u}^{-1}\mathrm{Q}\mathrm{A}_{\ell,u}^{-1}.$$

Thus clipping does not change the first-order influence function when the transfer conditions continue to hold. The plug-in matrix (10) *is valid exactly when*

$$\mathrm{A}_{\ell,u}^{-1}\mathrm{Q}\mathrm{A}_{\ell,u}^{-1} = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}.$$

If Q is positive definite, this equality holds if and only if $\mathrm{A} \in \mathcal{S}_{\ell,u}$. For singular Q, the displayed covariance equality is the exact validity condition and need not imply $\mathrm{A}_{\ell,u} = \mathrm{A}$. For a fixed contrast c, equality of the two quadratic forms in direction c is sufficient even when the full covariance matrices differ.

This result is conditional on the iterate rate and the other transfer conditions. A clipping box that makes the early recursion unstable can violate those conditions; projection consistency alone does not prevent that failure.

Remark C.25 (Raw second moments and smoothing bias)**.** *The estimator* (9) *averages raw score outer products. If $Y_t^\star = \zeta_t^\star + b_t(\theta_\star)$, then*

$$\mathbb{E}(Y_t^\star Y_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1}) = \mathbb{E}(\zeta_t^\star\zeta_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1}) + b_t(\theta_\star)b_t(\theta_\star)^\mathsf{T}.$$

Consequently, (42) *follows from the target quadratic-variation condition whenever $\frac{1}{n}\sum_{t=1}^{n}\|b_t(\theta_\star)\|^2 \longrightarrow 0$. This includes the centered and critical polynomial-bias regimes considered above. A fourth moment is not needed for this result. The argument does not cover a persistent approximation bias. Such a bias can change the target and its derivative, which requires a separate analysis of the modified mean score.*

Corollary C.26. *Under the centered transfer theorem:*

(i) $\mathrm{P}_t = \mathrm{I}_d$ *and a direct unbiased score give the classical Polyak-Ruppert averaged-SGD limit.*

(ii) Any fixed positive definite preconditioner satisfying the stability assumptions gives the same averaged influence function $-\mathrm{A}^{-1}G_\star$; $\mathrm{P} = \mathrm{A}^{-1}$ is oracle Newton preconditioning when A is symmetric positive definite.

(iii) A stochastic Newton or conditioned-gradient matrix is covered whenever it satisfies the dynamic stabilization condition.

(iv) $\mathrm{P}_t = \mathrm{I}_d$ *and a zeroth-order primary score reduce to an averaged Kiefer-Wolfowitz procedure with the covariance established in Chen et al. (2024).*

(v) A pathwise Lipschitz auxiliary score has $q = 0$ *by Proposition C.6; a fixed auxiliary batch therefore meets the curvature threshold for every* $\alpha \in (1/2, 1)$.

Remark C.27 (Other online covariance estimators)**.** *Batch means Chen et al. (2020), iterate-only online covariance estimation Zhu et al. (2023), its nonsmooth extension Jiang et al. (2025), and random scaling pivots Lee et al. (2022) avoid explicit estimation of both* A *and* Q*. The sandwich estimator is natural in the present method when the clipping box contains* A*, because the same curvature recursion then supplies its plug-in estimate.*

C.2.7 Comparison with SGD and stochastic Newton methods

The transfer theorem implies first-order equivalence among all preconditioners that satisfy the stabilization condition. The possible advantage of curvature learning is therefore a conditioning effect along the path, not a smaller limiting covariance. The following elementary calculation makes this distinction explicit.

Proposition C.28. *Let* A *be symmetric positive definite and let* $\Gamma = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$.

(i) Under the centered version of Theorem 3.2, every fixed positive definite preconditioner and every predictable dynamic preconditioner satisfying (T2) have the same limit $\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N(0, \Gamma)$.

(ii) Consider the noiseless linearized recursion $e_t = (\mathrm{I}_d - \eta_t \mathrm{P}\mathrm{A})e_{t-1}$ *with a fixed symmetric positive definite* P*. Put* $\mathrm{C}_\mathrm{P} = \mathrm{A}^{1/2}\mathrm{P}\mathrm{A}^{1/2}$ *and* $\|x\|_\mathrm{A} = (x^\mathsf{T}\mathrm{A}x)^{1/2}$*. If* $0 < \eta_t \lambda_{\max}(\mathrm{C}_\mathrm{P}) \le 1$*, then*

$$\|e_t\|_\mathrm{A} \le \exp\left\{-\lambda_{\min}(\mathrm{C}_\mathrm{P}) \sum_{s=1}^{t} \eta_s\right\} \|e_0\|_\mathrm{A} . \tag{85}$$

If $\eta_t = \rho \lambda_{\max}(\mathrm{C}_\mathrm{P})^{-1} t^{-\alpha}$ *with* $0 < \rho \le 1$*, the exponent in* (85) *is*

$$\frac{\rho}{\kappa(\mathrm{C}_\mathrm{P})} \sum_{s=1}^{t} s^{-\alpha}, \qquad \kappa(\mathrm{C}_\mathrm{P}) = \frac{\lambda_{\max}(\mathrm{C}_\mathrm{P})}{\lambda_{\min}(\mathrm{C}_\mathrm{P})}.$$

Oracle Newton preconditioning, $\mathrm{P} = \mathrm{A}^{-1}$*, gives* $\mathrm{C}_\mathrm{P} = \mathrm{I}_d$ *and condition number one.*

Part (ii) is a local deterministic statement, not a finite-sample dominance theorem. In the stochastic recursion, curvature noise, auxiliary evaluations, clipping, and factorization cost can offset the improved local condition number. It nevertheless identifies the setting in which the additional work is most plausible: A is ill conditioned, d is moderate, score evaluations are cheaper than exact Hessians, and early or continuously reported estimates matter.

The proposed method has three concrete advantages over unpreconditioned averaged SGD. First, when B_t approaches A, the local drift $\mathrm{B}_t^{-1}\mathrm{A}$ approaches the identity, reducing anisotropy in the recursion. Second, the curvature update uses score evaluations and remains meaningful for Lipschitz or discontinuous sample scores for which ordinary sample Hessians are unavailable. Third, the same matrix B_t is available for sandwich inference, so optimization and covariance estimation share a curvature estimate. These advantages do not imply uniform superiority. When A is well conditioned, exact Hessians are cheap, auxiliary scores are expensive, or d is too large for a dense matrix, averaged SGD or a structured stochastic Newton method can be preferable.

C.3 Functional limits and self-normalized inference

The pointwise central limit theorem is sufficient for plug-in Wald inference. Self-normalization instead uses the complete sequence of partial averages and therefore requires a process limit. For the single time-indexed score sequence considered here, the martingale variance condition already implies its uniform process counterpart. The additional requirements concern the maximal Abel

Table 4: Comparison with representative online methods. The covariance column concerns the centered fixed-dimensional Polyak-Ruppert limit under the regularity conditions of the corresponding method.

Method	Curvature input	Typical stored state	First-order inferential limit	Main tradeoff
Averaged SGD	none	$O(d)$	Gaussian PR limit with covariance $\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$	lowest cost; transient behavior can be poor under severe ill conditioning
Fixed or oracle-conditioned SGD	fixed P or known A^{-1}	$O(d^2)$ for dense P	the same Gaussian PR limit	improved local conditioning when $\mathrm{P} \approx \mathrm{A}^{-1}$; oracle curvature is usually unavailable
Stochastic Newton or quasi-Newton	sample Hessians, Hessian sketches, or smooth gradient differences	method dependent, often $O(d^2)$	usually the same Gaussian PR limit when the matrix stabilizes	can exploit exact smooth curvature; theory and cost depend on the matrix update
Score-secant recursion	paired auxiliary scores and random directions	$O(d^2)$ in the dense form	the same Gaussian PR limit under (28)	applies when a sample Hessian is unavailable; requires auxiliary queries and control of q, δ_t, R_t
Iterate-only random scaling	no curvature for inference	$O(d^2)$ for a vector pivot, or $O(1)$ per contrast	the same underlying PR process with a pivotal non-Gaussian studentized limit	avoids sandwich estimation but does not by itself improve conditioning of the primary recursion

boundary and the partial-sum behavior of target bias. In the centered direct score and undersmoothed regimes, the bias condition is automatic.

For $r \in [0,1]$, define

$$\mathbb{Z}_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} e_{t-1},$$

$$\mathbb{U}_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} \zeta_t^\star, \qquad \mathbb{B}_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} b_t(\theta_\star).$$

Thus $\mathbb{Z}_n(1) = \sqrt{n}(\bar{\theta}_n - \theta_\star)$.

For the fixed time-indexed covariance sequence, (T6) implies

$$\sup_{0 \le r \le 1} \left\| n^{-1} \sum_{t=1}^{\lfloor nr \rfloor} \mathbb{E}(\zeta_t^\star \zeta_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1}) - r\mathrm{Q} \right\|_{\mathrm{op}} \xrightarrow{p} 0. \tag{86}$$

Lemma C.40 proves this by monotonicity of positive semidefinite partial sums. The implication does not cover arbitrary horizon-dependent arrays.

Assumption C.29. *In addition to Assumption 3.1, the following conditions hold.*

(F2) For a deterministic continuous function $\mathbb{B} : [0,1] \to \mathbb{R}^d$, $\sup_{0 \le r \le 1} \|\mathbb{B}_n(r) - \mathbb{B}(r)\| \xrightarrow{p} 0$.

(F3) With $a_t = \eta_t^{-1}$,

$$\frac{1}{\sqrt{n}} \max_{1 \le m \le n} a_m \|\mathrm{M}_m e_m\| \xrightarrow{p} 0. \tag{87}$$

Condition (F2) is automatic with $\mathbb{B} \equiv 0$ when $n^{-1/2} \sum_{t=1}^n \|b_t(\theta_\star)\| \longrightarrow 0$. Condition (F3) is the uniform counterpart of the terminal boundary term in the pointwise Abel decomposition. No additional derivative of g, mixing assumption, or pointwise Hölder remainder is required at the level of the transfer theorem.

Theorem C.30 (Functional transfer theorem)**.** *Under Assumptions 3.1 and C.29, in* $D([0,1], \mathbb{R}^d)$ *equipped with the Skorokhod* J_1 *topology,*

$$\mathbb{Z}_n(\cdot) \Longrightarrow -\mathrm{A}^{-1}\mathrm{Q}^{1/2} W_d(\cdot) - \mathrm{A}^{-1}\mathbb{B}(\cdot), \tag{88}$$

where W_d is a standard d-dimensional Brownian motion. In particular, if $\mathbb{B} \equiv 0$ and $\Gamma = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$, then

$$\mathbb{Z}_n(\cdot) \Longrightarrow \Gamma^{1/2} W_d(\cdot), \tag{89}$$

where the sign in (88) *has been absorbed into the Brownian motion.*

The next proposition gives two transparent routes to the only genuinely new boundary condition.

Proposition C.31. *Suppose $a_t = \eta_0^{-1} t^{\alpha}$.*

(i) If, for a deterministic sequence $L_n = o\{n^{(1-\alpha)/2}\}$, $\max_{1 \le t \le n} t^{\alpha/2} \left\| \mathrm{M}_t e_t \right\| = O_p(L_n)$, then (87) *holds.*

(ii) If, for some $p > 2/(1-\alpha)$,

$$\sup_{t \ge 1} t^{\alpha p/2} \mathbb{E} \left\| \mathrm{M}_t e_t \right\|^p < \infty, \tag{90}$$

then (87) *holds.*

For the spectrally clipped recursion, $\sup_t \left\| \mathrm{M}_t \right\|_{\mathrm{op}} < \infty$ almost surely, so either condition can be verified from the corresponding iterate bound alone.

Related random scaling analyses use high noise moments together with additional regularity to control SGD paths (Lee et al., 2022). Here (90) is a sufficient iterate-moment condition, not an identification with those noise assumptions. It is not imposed by Theorem C.30: a sharper model-specific maximal inequality can establish (F3) under weaker moments.

Corollary C.32. *Consider the direct score versions of least squares, logistic, Huber, or quantile regression in Corollaries C.10 to C.12 and Theorem 3.5. Suppose their stated conditions hold and* (87) *is verified. Then* (89) *holds with the covariance matrix stated in the corresponding result.*

C.3.1 Self-normalized confidence sets

Define the recursive-path normalizer

$$\mathrm{V}_n = \frac{1}{n^2} \sum_{s=1}^{n} s^2 (\bar{\theta}_s - \bar{\theta}_n)(\bar{\theta}_s - \bar{\theta}_n)^{\mathsf{T}}.$$

The identity

$$\mathrm{V}_n = \frac{1}{n} \sum_{s=1}^{n} \left\{ \mathbb{Z}_n(s/n) - \frac{s}{n} \mathbb{Z}_n(1) \right\} \left\{ \mathbb{Z}_n(s/n) - \frac{s}{n} \mathbb{Z}_n(1) \right\}^{\mathsf{T}} \tag{91}$$

shows why a functional limit is the natural input. Let

$$\overline{W}_\ell(r) = W_\ell(r) - r W_\ell(1), \qquad \mathrm{J}_\ell = \int_0^1 \overline{W}_\ell(r) \overline{W}_\ell(r)^{\mathsf{T}} \, dr.$$

The random matrix J_ℓ is positive definite almost surely.

Theorem C.33. *Suppose the centered functional limit* (89) *holds and let $\mathrm{R} \in \mathbb{R}^{\ell \times d}$ have rank ℓ, with $\mathrm{R}\Gamma\mathrm{R}^{\mathsf{T}}$ positive definite. Then $\mathrm{R}\mathrm{V}_n\mathrm{R}^{\mathsf{T}}$ is nonsingular with probability tending to one. Under $H_0 : \mathrm{R}\theta_\star = r_0$,*

$$T_n^{\mathrm{SN}} := n(\mathrm{R}\bar{\theta}_n - r_0)^{\mathsf{T}} (\mathrm{R}\mathrm{V}_n\mathrm{R}^{\mathsf{T}})^{-1} (\mathrm{R}\bar{\theta}_n - r_0) \Longrightarrow W_\ell(1)^{\mathsf{T}} \mathrm{J}_\ell^{-1} W_\ell(1). \tag{92}$$

Consequently, if $q^{\mathrm{SN}}_{\ell,1-\gamma}$ is the $(1-\gamma)$ quantile of the limit in (92)*, then*

$$\mathcal{C}^{\mathrm{SN}}_{n,1-\gamma} = \left\{ r \in \mathbb{R}^\ell : n(\mathrm{R}\bar{\theta}_n - r)^{\mathsf{T}} (\mathrm{R}\mathrm{V}_n\mathrm{R}^{\mathsf{T}})^{-1} (\mathrm{R}\bar{\theta}_n - r) \le q^{\mathrm{SN}}_{\ell,1-\gamma} \right\}$$

satisfies $\mathbb{P}\{\mathrm{R}\theta_\star \in \mathcal{C}^{\mathrm{SN}}_{n,1-\gamma}\} \to 1-\gamma$. On the event that $\mathrm{R}\mathrm{V}_n\mathrm{R}^{\mathsf{T}}$ is singular, define the statistic as zero and the region as $\mathbb{R}^\ell$; that event has probability tending to zero.

For a contrast $c \in \mathbb{R}^d$ satisfying $c^{\mathsf{T}}\Gamma c > 0$, let $q^{\mathrm{sn}}_{1-\gamma/2}$ be the $(1-\gamma/2)$ quantile of

$$\frac{W_1(1)}{\left\{ \int_0^1 [W_1(r) - r W_1(1)]^2 \, dr \right\}^{1/2}},$$

then

$$c^{\mathsf{T}}\bar{\theta}_n \;\pm\; q^{\mathrm{sn}}_{1-\gamma/2}\sqrt{\frac{c^{\mathsf{T}}\mathrm{V}_n c}{n}} \tag{93}$$

has asymptotic coverage $1-\gamma$ *for* $c^{\mathsf{T}}\theta_\star$.

The normalizer is online. Let

$$C_n^{(0)} = \sum_{s=1}^{n} s^2, \qquad C_n^{(1)} = \sum_{s=1}^{n} s^2\bar{\theta}_s, \qquad \mathrm{C}_n^{(2)} = \sum_{s=1}^{n} s^2\bar{\theta}_s\bar{\theta}_s^{\mathsf{T}}.$$

These quantities satisfy one-step recursions, and

$$\mathrm{V}_n = n^{-2}\left\{\mathrm{C}_n^{(2)} - C_n^{(1)}\bar{\theta}_n^{\mathsf{T}} - \bar{\theta}_n C_n^{(1)\mathsf{T}} + C_n^{(0)}\bar{\theta}_n\bar{\theta}_n^{\mathsf{T}}\right\}.$$

Thus vector self-normalization uses $O(d^2)$ memory and no resampling; a fixed scalar contrast uses $O(1)$ additional memory. The critical values depend only on Brownian motion and can be simulated once.

Relation to existing self-normalization. The construction follows recursive-estimate self-normalization (Shao, 2010a;b) and random scaling for averaged SGD (Lee et al., 2022). The new step here is uniform transfer through the changing Newton matrix, rather than a new pivot. Self-normalization avoids consistent estimation of the asymptotic covariance, although it still computes a random quadratic normalizer. This can be useful when nuisance estimation is difficult: Liu et al. (2023) use it for population-quantile inference under local differential privacy. That procedure includes a privacy mechanism absent from the present recursion. Since our method already learns curvature, sandwich inference is a natural default; the numerical appendix compares its calibration with self-normalization as the sample size grows.

Bias. The pivotal conclusion requires the centered process limit. If $\mathbb{B}$ is nonzero, the numerator and generally the normalizer contain an unknown deterministic drift. For the critical zeroth-order regime in Theorem C.18, $b_t(\theta_\star) \sim h_0^p t^{-1/2} B$ and the integral bounds for $\sum_{t\le nr} t^{-1/2}$ give, uniformly in $r \in [0,1]$, $\mathbb{B}(r) = 2h_0^p\sqrt{r}\,B$, whose bridge drift is $2h_0^p(\sqrt{r}-r)B$ and is not removed by the centering in (91). Self-normalized inference is therefore valid without further correction in the direct score and undersmoothed regimes. A critical-rate implementation needs a consistent pathwise bias correction, not only an endpoint recentering.

C.4 Assumptions and model-specific verification

The transfer conditions separate stability of the parameter path from the additional work needed to learn curvature. The mean-square rate (T1) controls the boundary and nonlinear terms. It is the usual scale under strong convexity and suitable stepsizes (Polyak & Juditsky, 1992; Moulines & Bach, 2011; Kushner & Yin, 2003); Proposition C.42 gives a sufficient proof on a valid stability region. Local positive curvature alone does not establish this region or prevent early instability. The scalar constructions in section C.5.9 show that the stated conditions are attainable for logistic, quantile and Huber scores.

The changing preconditioner introduces the Abel term in (T2). The stabilization threshold is due to An & Huo (2026); for score secants, Theorem 3.3 derives it from the second moment exponent, radius and auxiliary batch size. The projection set need only contain A, not contain it in its strict interior. Without membership, the recursion learns $\mathrm{A}_{\ell,u}$ and the plug-in sandwich is generally inconsistent, as shown in Corollary C.24. This distinction explains why a stable clipped update and a valid variance estimate are different requirements.

Conditions (T3)–(T5) are normalized remainders rather than stronger pointwise smoothness requirements. A Hölder population Jacobian is one sufficient route for (T3). Quadratic-mean score continuity verifies (T4), with exponent 2 for Lipschitz sample scores and exponent 1 for quantile indicators under a density envelope. Direct unbiased scores make (T5) automatic; finite difference scores require control of local bias variation, for example through Proposition C.17. The conditional variance and Lindeberg limits in (T6) are standard martingale conditions (Hall & Heyde, 1980). An iid target score with a $2+\epsilon$ moment is a convenient sufficient case.

The secant assumptions concern the population derivative and the score increments separately. In (S1), continuity of Dg on the compact perturbation tube yields uniform continuity; no derivative of the Jacobian or sample Hessian is required. Bounded isotropic directions in (S2) keep perturbations inside that tube. Gaussian directions require global bounds or integrable domination instead. The exponent in (S3) is calculated from the sample score: pathwise Lipschitz scores give $q = 0$, shared quantile endpoints give $q = 1$, and independent noisy endpoints give $q = 2$. Independent streams supply the conditional moments directly. Proposition C.8 also permits a current primary observation when both secants are centered at θ_{t-1} and their matrix is first used at the next update.

In the smooth regression specializations, the score Lipschitz envelope is proportional to $\|X\|^2$, so a fourth design moment supplies a sufficient raw-secant second moment bound. It is not a general requirement of the sandwich central limit theorem. Least squares also uses a $2 + \epsilon$ moment of εX for Lindeberg and the outer-product law. For logistic and Huber regression, global score bounds control excursions from the neighborhood, but the iterate rate remains a separate input. Huber regression needs a local density envelope at the two clipping thresholds, not a density derivative (Huber, 1973).

For quantile regression, (Q1)–(Q2) control the population expansion, discontinuity-crossing probability and target-score moments (Koenker, 2005). Their integrated envelopes allow unbounded covariates and do not impose a fourth design moment. Positive-definite sensitivity in (Q3) gives local identification. The sampling requirement in (Q4), or the explicitly permitted reuse construction, supplies conditional moments. Localization and the iterate rate in (Q5) must hold for the actual trajectory; projecting the parameter would require a rate analysis of that modified algorithm.

The finite difference extension uses ordinary differentiability of the population score for the centered bias bound in Proposition C.20. The sharper expansion in Proposition C.21 is needed only to identify the critical mean shift. Separately, recursive sandwich consistency requires the local second moment replacement and target outer-product law in Theorem 3.6; a uniform $2 + \epsilon$ target moment suffices, without a fourth score moment.

For functional inference, (T6) already gives uniform quadratic variation for the fixed time-indexed sequence, by Lemma C.40. The additional conditions are the partial-sum bias limit (F2) and maximal Abel boundary (F3). Direct unbiased scores and uniform undersmoothing make the bias path zero. Proposition C.31 gives sufficient maximal-rate and higher-moment routes for the boundary. An uncorrected critical bias remains in the self-normalizer and is outside the centered pivotal conclusion.

C.5 Proofs

C.5.1 Proofs for the transfer theorem

An algebraic summation identity

Lemma C.34 (Matrix summation by parts)**.** *For $a_t = \eta_t^{-1}$, arbitrary matrices M_t, and vectors e_t,*

$$\sum_{t=1}^{n} a_t \mathrm{M}_t (e_{t-1} - e_t) = a_1 \mathrm{M}_1 e_0 - a_n \mathrm{M}_n e_n + \sum_{t=1}^{n-1} (a_{t+1}\mathrm{M}_{t+1} - a_t \mathrm{M}_t) e_t. \tag{94}$$

Proof. Expand the left side as

$$\sum_{t=1}^{n} a_t \mathrm{M}_t e_{t-1} - \sum_{t=1}^{n} a_t \mathrm{M}_t e_t.$$

In the first sum, isolate $t = 1$ and reindex the remaining terms by $s = t - 1$:

$$\sum_{t=1}^{n} a_t \mathrm{M}_t e_{t-1} = a_1 \mathrm{M}_1 e_0 + \sum_{s=1}^{n-1} a_{s+1} \mathrm{M}_{s+1} e_s.$$

In the second sum, isolate $t = n$:

$$\sum_{t=1}^{n} a_t \mathrm{M}_t e_t = \sum_{t=1}^{n-1} a_t \mathrm{M}_t e_t + a_n \mathrm{M}_n e_n.$$

Subtracting the last two displays and combining the coefficients of e_t, $1 \le t \le n-1$, gives (94). □

The elementary identity is included only to make the later decomposition transparent.

Dynamic-remainder bound

Proof of Proposition C.1. Write $\|X\|_{L^2(\mathrm{op})} = \{\mathbb{E}\,\|X\|_{\mathrm{op}}^2\}^{1/2}$. Because $a_t = \eta_0^{-1} t^\alpha$, the mean-value theorem gives

$$a_{t+1} - a_t \le C t^{\alpha-1}. \tag{95}$$

For $t \ge 1$,

$$a_{t+1}\mathrm{M}_{t+1} - a_t\mathrm{M}_t = (a_{t+1} - a_t)\mathrm{M}_{t+1} + a_t(\mathrm{M}_{t+1} - \mathrm{M}_t),$$

so (14) and (95) imply

$$\|a_{t+1}\mathrm{M}_{t+1} - a_t\mathrm{M}_t\|_{L^2(\mathrm{op})} \le C\{t^{\alpha-1} + t^{\alpha-\beta_{\mathrm{P}}}\}.$$

Cauchy-Schwarz, (13), and the uniform L^2 bound on M_t give

$$\frac{1}{\sqrt{n}}\mathbb{E}\,\|a_n\mathrm{M}_n e_n\| \le \frac{a_n}{\sqrt{n}}\,\|\mathrm{M}_n\|_{L^2(\mathrm{op})}\,\{\mathbb{E}\,\|e_n\|^2\}^{1/2} \le C n^{(\alpha-1)/2} \longrightarrow 0.$$

The initial boundary term is $O(n^{-1/2})$ by the same inequality and $\mathbb{E}\,\|e_0\|^2 < \infty$. For the increment sum, another application of Cauchy-Schwarz yields

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1}^{n-1}\mathbb{E}\,\|(a_{t+1}\mathrm{M}_{t+1} - a_t\mathrm{M}_t)e_t\| \\
&\quad\le \frac{C}{\sqrt{n}}\sum_{t=1}^{n-1}\{t^{\alpha-1} + t^{\alpha-\beta_{\mathrm{P}}}\}t^{-\alpha/2} \\
&\quad= \frac{C}{\sqrt{n}}\sum_{t=1}^{n-1}\{t^{\alpha/2-1} + t^{\alpha/2-\beta_{\mathrm{P}}}\}.
\end{aligned} \tag{96}$$

The first sum in (96) is $O(n^{(\alpha-1)/2})$. The second is

$$O\Big(n^{(1+\alpha)/2-\beta_{\mathrm{P}}}\Big) + O(n^{-1/2}\log n) + O(n^{-1/2}), \tag{97}$$

where the latter two terms cover the boundary and summable exponent cases. It tends to zero because $\beta_{\mathrm{P}} > (1+\alpha)/2$. Combining the two boundary terms with (96) proves $\mathbb{E}\left\|\mathcal{R}_n^{\mathrm{dyn}}\right\| \to 0$. Markov's inequality gives convergence in probability. □

Transfer-theorem proof

Proof of Theorem 3.2. For each t, combine (6), (1), (2), and (12):

$$e_{t-1} - e_t = \eta_t \mathrm{P}_t\{\mathrm{A}e_{t-1} + r(e_{t-1}) + b_t(\theta_{t-1}) + \zeta_t(\theta_{t-1})\}. \tag{98}$$

Multiply (98) on the left by $\eta_t^{-1}\mathrm{M}_t$. From $\mathrm{M}_t = (\mathrm{P}_t\mathrm{A})^{-1}$, $\mathrm{M}_t\mathrm{P}_t\mathrm{A} = \mathrm{I}_d$, $\quad \mathrm{M}_t\mathrm{P}_t = \mathrm{A}^{-1}$. Therefore

$$\begin{aligned}
e_{t-1} = \eta_t^{-1}\mathrm{M}_t(e_{t-1} - e_t) &- \mathrm{A}^{-1}\zeta_t^\star - \mathrm{A}^{-1}\{\zeta_t(\theta_{t-1}) - \zeta_t^\star\} \\
&- \mathrm{A}^{-1}r(e_{t-1}) - \mathrm{A}^{-1}b_t(\theta_\star) - \mathrm{A}^{-1}\{b_t(\theta_{t-1}) - b_t(\theta_\star)\}.
\end{aligned} \tag{99}$$

Sum (99) over $t = 1, \dots, n$ and divide by $\sqrt{n}$. We treat its six terms in order.

First, apply Lemma C.34 with $a_t = \eta_t^{-1}$. The first normalized sum in (99) is exactly $\mathcal{R}_n^{\mathrm{dyn}}$ in (44). Hence Proposition C.1 gives

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\eta_t^{-1}\mathrm{M}_t(e_{t-1} - e_t) = \mathcal{R}_n^{\mathrm{dyn}} = o_p(1).$$

Second, define $\Delta_t = \zeta_t(\theta_{t-1}) - \zeta_t^\star$. Both terms in the difference have conditional mean zero, so $\mathbb{E}(\Delta_t \mid \mathcal{F}_{t-1}) = 0$. Martingale orthogonality yields

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\Delta_t\right\|^2 = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\,\|\Delta_t\|^2$$

$$\longrightarrow 0$$

by (16). Hence the local score-replacement term is $o_p(1)$.

Third, Markov's inequality and (15) give $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} r(e_{t-1}) = o_p(1)$. Fourth, the same argument with (17) gives $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{b_t(\theta_{t-1}) - b_t(\theta_\star)\} = o_p(1)$. The remaining two terms are the target martingale sum and the target bias sum. We have proved

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} e_{t-1} = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\zeta_t^\star - \mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} b_t(\theta_\star) + o_p(1). \tag{100}$$

By definition (8), the left side of (100) equals $\sqrt{n}(\bar{\theta}_n - \theta_\star)$, proving (21).

Conditions (18)–(19) give, by the martingale central limit theorem, $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\zeta_t^\star \implies N(0, \mathrm{Q})$. Combine this convergence with (20), (21), and Slutsky's theorem. A linear transformation of the Gaussian limit has covariance $\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$, which proves (22). □

Primitive-condition proof

Proof of Proposition C.3. Let $t_1 = 1$ when the primitive inequalities hold globally and $t_1 = t_0$ under the local alternative. The finite prefix contributes $O(n^{-1/2})$ to (15) and (17), and $O(n^{-1})$ to (16). It therefore suffices to control the sums from t_1 onward; when evaluating the deterministic rate sums, replacing a fixed lower limit by one does not change the rate.

Because $1 + \nu \le 2$, Lyapunov's inequality and (13) imply

$$\mathbb{E}\,\|e_{t-1}\|^{1+\nu} \le \{\mathbb{E}\,\|e_{t-1}\|^2\}^{(1+\nu)/2} \le C\eta_{t-1}^{(1+\nu)/2} \le Ct^{-\alpha(1+\nu)/2}.$$

Using (45),

$$\frac{1}{\sqrt{n}}\sum_{t=t_1}^{n}\mathbb{E}\,\|r(e_{t-1})\| \le \frac{C}{\sqrt{n}}\sum_{t=t_1}^{n} t^{-\alpha(1+\nu)/2}.$$

For $a = \alpha(1+\nu)/2$, the last expression is $O(n^{1/2-a})$ when $a < 1$, $O(n^{-1/2}\log n)$ when $a = 1$, and $O(n^{-1/2})$ when $a > 1$. It tends to zero exactly under $a > 1/2$, which is (48). This proves (15).

Similarly, because $\chi \le 2$,

$$\mathbb{E}\,\|e_{t-1}\|^{\chi} \le \{\mathbb{E}\,\|e_{t-1}\|^2\}^{\chi/2} \le Ct^{-\alpha\chi/2}.$$

Taking expectations in (46) and averaging gives

$$\frac{1}{n}\sum_{t=t_1}^{n}\mathbb{E}\,\|\zeta_t(\theta_{t-1}) - \zeta_t(\theta_\star)\|^2 \le \frac{C}{n}\sum_{t=t_1}^{n} t^{-\alpha\chi/2} + \frac{1}{n}\sum_{t=t_1}^{n}\rho_t.$$

The first term tends to zero for every $\chi > 0$, and the second tends to zero by assumption. This proves (16).

Finally, (47), Lyapunov's inequality, and (13) give

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{t=t_1}^{n}\mathbb{E}\,\|b_t(\theta_{t-1}) - b_t(\theta_\star)\| &\le \frac{1}{\sqrt{n}}\sum_{t=t_1}^{n} L_t\{\mathbb{E}\,\|e_{t-1}\|^2\}^{\xi/2} \\ &\le \frac{C}{\sqrt{n}}\sum_{t=t_1}^{n} L_t\eta_{t-1}^{\xi/2} \longrightarrow 0.\end{aligned}$$

This is (17). □

C.5.2 Proofs for the score secant recursion

Generic score secant moments

Proof of Proposition C.5. Fix $\theta \in \mathcal{U}$, $0 < \delta \le \bar{\delta}$, and one direction $V = v$. By (3),

$$\mathbb{E}_W\left[\frac{\Psi(\theta+\delta v; W) - \Psi(\theta-\delta v; W)}{2\delta}\right] = \frac{g(\theta+\delta v) - g(\theta-\delta v)}{2\delta}.$$

The fundamental theorem of calculus along the line segment gives

$$\frac{g(\theta+\delta v) - g(\theta-\delta v)}{2\delta} = \frac{1}{2}\int_{-1}^{1} \mathrm{H}(\theta + s\delta v)v\, ds. \tag{101}$$

After multiplication by v^{T}, symmetrization, and expectation over V,

$$\mathbb{E}\mathrm{K}_\delta(\theta; W, V) = \frac{1}{2}\int_{-1}^{1} \mathbb{E}_V\{\mathrm{sym}(\mathrm{H}(\theta + s\delta V)VV^{\mathsf{T}})\}\, ds.$$

Since $\mathrm{H}(\theta)$ is symmetric and $\mathbb{E}(VV^{\mathsf{T}}) = \mathrm{I}_d$, $\mathbb{E}_V\{\mathrm{sym}(\mathrm{H}(\theta)VV^{\mathsf{T}})\} = \mathrm{H}(\theta)$. The tube condition, $\|V\| \le \bar{v}$, and the definition of ω_{H} imply

$$\begin{aligned}\left\|\mathbb{E}\mathrm{K}_\delta(\theta; W, V) - \mathrm{H}(\theta)\right\|_{\mathrm{F}} &\le \frac{1}{2}\int_{-1}^{1} \mathbb{E}\{\left\|\mathrm{H}(\theta + s\delta V) - \mathrm{H}(\theta)\right\|_{\mathrm{F}} \|V\|^2\}\, ds \\ &\le \bar{v}^2\omega_{\mathrm{H}}(\bar{v}\delta).\end{aligned}$$

Also $\left\|\mathrm{H}(\theta) - \mathrm{A}\right\|_{\mathrm{F}} \le \omega_{\mathrm{H}}(\|\theta - \theta_\star\|)$. Combining these bounds and taking $\theta = \theta_{t-1}$ on $\{\theta_{t-1} \in \mathcal{U}\}$ proves (50).

For the conditional variance, first condition on V_t. Independence of the R_t observations gives

$$\mathbb{E}\left[\left\|\widehat{\mathrm{H}}_t - \mathbb{E}(\widehat{\mathrm{H}}_t \mid V_t, \mathcal{F}_{t-1})\right\|_{\mathrm{F}}^2 \mid V_t, \mathcal{F}_{t-1}\right] \le \frac{1}{R_t}\mathbb{E}(\left\|\mathrm{sym}(D_{t1}V_t^{\mathsf{T}})\right\|_{\mathrm{F}}^2 \mid V_t, \mathcal{F}_{t-1}).$$

For vectors a, v, $\left\|\mathrm{sym}(av^{\mathsf{T}})\right\|_{\mathrm{F}} \le \|a\|\,\|v\|$. Therefore

$$\left\|\mathrm{sym}(D_{t1}V_t^{\mathsf{T}})\right\|_{\mathrm{F}}^2 \le \frac{1}{4\delta_t^2}\left\|\Psi(\theta_{t-1} + \delta_t V_t; W_{t1}) - \Psi(\theta_{t-1} - \delta_t V_t; W_{t1})\right\|^2 \|V_t\|^2.$$

Condition (S3) yields $\mathbb{E}\left\|\mathrm{sym}(D_{t1}V_t^{\mathsf{T}})\right\|_{\mathrm{F}}^2 \le C\delta_t^{-q}$, so the sampling component is at most $C/(R_t\delta_t^q)$. Continuity of H on a compact set and boundedness of V make the conditional mean in (101) uniformly bounded. Its variance with respect to V is therefore $O(1)$. The law of total variance proves (51). □

Proof of Proposition C.6. By (52), $\|\Psi(\theta+\delta V; W) - \Psi(\theta-\delta V; W)\| \le 2\delta L(W)\|V\|$. Consequently,

$$\mathbb{E}\Big[\left\|\Psi(\theta+\delta V; W) - \Psi(\theta-\delta V; W)\right\|^2 \|V\|^2\Big] \le 4\delta^2\mathbb{E}\{L(W)^2\|V\|^4\} \le C\delta^2,$$

which is (49) with $q = 0$. □

Projection, consistency, and increments

Lemma C.35 (Projection increment bound)**.** *Under* (24)*,* (25)*, and* (23)*,*

$$\{\mathbb{E}\left\|\mathrm{B}_t - \mathrm{B}_{t-1}\right\|_{\mathrm{F}}^2\}^{1/2} \le Ct^{-1+s_{\mathrm{H}}/2}.$$

Proof. Metric projection onto a closed convex set in a Hilbert space is nonexpansive. Since $\mathrm{B}_{t-1} \in \mathcal{S}_{\ell,u}$, $\left\|\mathrm{B}_t - \mathrm{B}_{t-1}\right\|_{\mathrm{F}} \le t^{-1}\left\|\widehat{\mathrm{H}}_t - \mathrm{B}_{t-1}\right\|_{\mathrm{F}}$. The set $\mathcal{S}_{\ell,u}$ is bounded. Condition (24) implies that $\mathbb{E}\left\|\mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1})\right\|_{\mathrm{F}}^2$ is bounded after a finite prefix. Together with (25), this gives

$$\mathbb{E}\left\|\widehat{\mathrm{H}}_t\right\|_{\mathrm{F}}^2 \le C\{1 + (R_t\delta_t^q)^{-1}\} \le C(1 + t^{s_{\mathrm{H}}}).$$

Therefore

$$\{\mathbb{E}\left\|\mathrm{B}_t - \mathrm{B}_{t-1}\right\|_{\mathrm{F}}^2\}^{1/2} \le Ct^{-1}(1 + t^{s_{\mathrm{H}}/2}) \le Ct^{-1+s_{\mathrm{H}}/2}.$$

□

Lemma C.36 (Projected secant restricted consistency)**.** *Let* A *be symmetric positive definite, let* $\mathrm{A}_{\ell,u} = \Pi_{\ell,u}(\mathrm{A})$*, and suppose* $s_{\mathrm{H}} < 1$*, the conditional variance bound* (25) *holds under the schedules in* (23)*, and* (24) *holds. Then* $\mathbb{E}\,\|\mathrm{B}_t - \mathrm{A}_{\ell,u}\|_{\mathrm{F}}^2 \to 0$*.*

Proof. Put $\mathrm{E}_t = \mathrm{B}_t - \mathrm{A}_{\ell,u}$ and $\mathcal{V}_t = \|\mathrm{E}_t\|_{\mathrm{F}}^2$. The characterization of metric projection gives

$$\langle \mathrm{A} - \mathrm{A}_{\ell,u}, \mathrm{B} - \mathrm{A}_{\ell,u} \rangle_{\mathrm{F}} \le 0 \qquad \text{for every } \mathrm{B} \in \mathcal{S}_{\ell,u}. \tag{102}$$

Since $\mathrm{A}_{\ell,u}$ belongs to the projection set and projection is nonexpansive,

$$\mathcal{V}_t \le \left\| (1 - t^{-1})\mathrm{E}_{t-1} + t^{-1}(\widehat{\mathrm{H}}_t - \mathrm{A}_{\ell,u}) \right\|_{\mathrm{F}}^2. \tag{103}$$

Let $\mathrm{D}_t = \mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1}) - \mathrm{A}$. Taking the conditional expectation of (103) gives

$$\begin{aligned}\mathbb{E}(\mathcal{V}_t \mid \mathcal{F}_{t-1}) \le (1 - t^{-1})^2 \mathcal{V}_{t-1} + 2t^{-1}(1 - t^{-1}) \langle \mathrm{E}_{t-1}, \mathrm{D}_t + \mathrm{A} - \mathrm{A}_{\ell,u} \rangle_{\mathrm{F}} \\ + t^{-2}\mathbb{E}(\left\| \widehat{\mathrm{H}}_t - \mathrm{A}_{\ell,u} \right\|_{\mathrm{F}}^2 \mid \mathcal{F}_{t-1}).\end{aligned}$$

By (102), $\langle \mathrm{E}_{t-1}, \mathrm{A} - \mathrm{A}_{\ell,u} \rangle_{\mathrm{F}} \le 0$. Because B_{t-1} and $\mathrm{A}_{\ell,u}$ lie in the bounded set $\mathcal{S}_{\ell,u}$, $\|\mathrm{E}_{t-1}\|_{\mathrm{F}} \le C$. Hence, after taking expectations, $\mathbb{E}\,|\langle \mathrm{E}_{t-1}, \mathrm{D}_t \rangle_{\mathrm{F}}| \le C d_{t,2}^{1/2}$. The second moment bound used in Lemma C.35 yields

$$t^{-2}\mathbb{E}\left\| \widehat{\mathrm{H}}_t - \mathrm{A}_{\ell,u} \right\|_{\mathrm{F}}^2 = O(t^{-2+s_{\mathrm{H}}}) = o(t^{-1})$$

because $s_{\mathrm{H}} < 1$. Thus, for $v_t = \mathbb{E}\mathcal{V}_t$,

$$v_t \le (1 - c/t)v_{t-1} + \frac{\varepsilon_t}{t}, \qquad \varepsilon_t \longrightarrow 0, \tag{104}$$

for all sufficiently large t and some $c > 0$. To see this, absorb the $o(t^{-1})$ term into ε_t/t and use $(1 - t^{-1})^2 \le 1 - c/t$.

Iterating (104) from a fixed t_0 gives

$$v_t \le v_{t_0} \prod_{j=t_0+1}^{t} (1 - c/j) + \sum_{k=t_0+1}^{t} \frac{\varepsilon_k}{k} \prod_{j=k+1}^{t} (1 - c/j).$$

For a constant C, the product is at most $C(k/t)^c$. The initial term tends to zero. Given $\varepsilon > 0$, choose K so that $|\varepsilon_k| \le \varepsilon$ for $k \ge K$. The finite sum over $k < K$ tends to zero, while the remaining tail is bounded by $C\varepsilon t^{-c} \sum_{k=K}^{t} k^{c-1} \le C'\varepsilon$. Taking the limsup and then $\varepsilon \downarrow 0$ proves $v_t \to 0$. □

Proof of Theorem 3.3. Lemma C.36 gives consistency in (26) under (24). The increment bound (27) follows from Lemma C.35.

With $\mathrm{P}_t = \mathrm{B}_{t-1}^{-1}$, $\mathrm{M}_t = (\mathrm{P}_t\mathrm{A})^{-1} = \mathrm{A}^{-1}\mathrm{B}_{t-1}$. Therefore

$$\begin{aligned}\{\mathbb{E}\,\|\mathrm{M}_t - \mathrm{M}_{t-1}\|_{\mathrm{op}}^2\}^{1/2} &\le \left\|\mathrm{A}^{-1}\right\|_{\mathrm{op}} \{\mathbb{E}\,\|\mathrm{B}_{t-1} - \mathrm{B}_{t-2}\|_{\mathrm{F}}^2\}^{1/2} \\ &\le C t^{-1+s_{\mathrm{H}}/2}.\end{aligned}$$

This is (14) with $\beta_{\mathrm{P}} = 1 - s_{\mathrm{H}}/2$. The inequality $\beta_{\mathrm{P}} > (1+\alpha)/2$ is equivalent to $s_{\mathrm{H}} < 1 - \alpha$. Since $1 - \alpha < 1$, this condition also implies $s_{\mathrm{H}} < 1$. Continuity of inversion on $\mathcal{S}_{\ell,u}$ gives $\mathrm{P}_t \xrightarrow{p} \mathrm{A}_{\ell,u}^{-1}$.

It remains to verify the last assertion. Under Assumption C.4, (50) gives

$$\left\| \mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1}) - \mathrm{A} \right\|_{\mathrm{F}} \le C\{\omega_{\mathrm{H}}(\|e_{t-1}\|) + \omega_{\mathrm{H}}(\bar{v}\delta_t)\}.$$

The right side is bounded because H is continuous on a compact set. If $e_{t-1} \xrightarrow{p} 0$, then the first term converges to zero in probability. Boundedness gives convergence in L^2, while the second term tends to zero deterministically. Thus $d_{t,2} \to 0$. □

Proof of Proposition C.8. At the start of iteration t, θ_{t-1}, B_{t-1}, and $\mathrm{P}_t = \mathrm{B}_{t-1}^{-1}$ are $\mathcal{F}_{t-1}$-measurable. Thus the primary update uses a predictable preconditioner. Conditional on $\mathcal{F}_{t-1}$, the collection $W_{t1} = Z_t, W_{t2}, \ldots, W_{tR_t}$ consists of independent observations with the auxiliary law and is independent of V_t. Because the center θ_{t-1} is also measurable, the conditional law used in the proof of Proposition C.5 is unchanged.

Let $\mathcal{F}_t$ augment $\mathcal{F}_{t-1}$ by the current primary observation, the remaining auxiliary observations, the current direction, and all quantities computed from them. Then B_t is $\mathcal{F}_t$-measurable and $\mathrm{P}_{t+1} = \mathrm{B}_t^{-1}$ is predictable at iteration $t+1$. The shared observation can make $\widehat{\mathrm{H}}_t$ and the current primary noise dependent, but none of Proposition C.5 and Theorems 3.2 and 3.3 assumes their within-iteration independence. The covariance proof likewise uses the outer-product conditions in Theorem 3.6, followed by convergence in probability of B_t; it does not require independence between B_t and the current outer product. This proves the claim. □

C.5.3 Verification for standard regression problems

Least squares

Proof of Corollary C.10. Let $e = \theta - \theta_{\mathrm{LS}}$. The population score is exactly linear: $g_{\mathrm{LS}}(\theta) = \mathbb{E}\{X(X^\mathsf{T}\theta - Y)\} = \mathrm{A}_{\mathrm{LS}}e$. Thus the mean remainder in (12) is zero. The finite quadratic Lyapunov function

$$\widetilde{R}_{\mathrm{LS}}(\theta) = \frac{1}{2}(\theta - \theta_{\mathrm{LS}})^\mathsf{T}\mathrm{A}_{\mathrm{LS}}(\theta - \theta_{\mathrm{LS}})$$

has gradient g_{LS} and is globally strongly convex and smooth. It replaces the uncentered expected squared loss, which need not be finite under (53).

For a fresh primary observation,

$$\begin{aligned}\zeta(\theta) &= X(X^\mathsf{T}\theta - Y) - \mathrm{A}_{\mathrm{LS}}e \\ &= (XX^\mathsf{T} - \mathrm{A}_{\mathrm{LS}})e - \varepsilon X.\end{aligned}$$

The fourth moment of X and the second moment implied by (53) give

$$\mathbb{E}\{\|\zeta(\theta)\|^2 \mid \mathcal{F}_{t-1}\} \le C(1 + \|e\|^2) \le C\{1 + \|g_{\mathrm{LS}}(\theta)\|^2\},$$

where nonsingularity of A_{LS} is used in the second inequality. Since spectral clipping gives $u^{-1}\mathrm{I}_d \preceq \mathrm{P}_t \preceq \ell^{-1}\mathrm{I}_d$, Proposition C.42 yields $\mathbb{E}\,\|e_t\|^2 = O(\eta_t)$ for sufficiently small η_0.

The centered local score difference satisfies $\zeta(\theta) - \zeta(\theta_{\mathrm{LS}}) = (XX^\mathsf{T} - \mathrm{A}_{\mathrm{LS}})e$, so $\mathbb{E}\,\|\zeta(\theta) - \zeta(\theta_{\mathrm{LS}})\|^2 \le C\,\|e\|^2$. This verifies (16). At the target, $\zeta^\star = -\varepsilon X$. The $(2+\epsilon_0)$ moment in (53) gives the Lindeberg condition, and the target covariance is Q_{LS}.

For the auxiliary score, $G_{\mathrm{LS}}(\theta + \delta V; X, Y) - G_{\mathrm{LS}}(\theta - \delta V; X, Y) = 2\delta XX^\mathsf{T}V$. Hence the secant does not depend on δ, and $\mathbb{E}_{X,V}\{\mathrm{sym}(XX^\mathsf{T}VV^\mathsf{T})\} = \mathbb{E}(XX^\mathsf{T}) = \mathrm{A}_{\mathrm{LS}}$. Its second moment is finite by $\mathbb{E}\,\|X\|^4 < \infty$, so $q = 0$, $d_{t,2} = 0$, and Theorem 3.3 gives the required stabilization. Applying Theorem 3.2 proves (54)–(55).

For covariance estimation,

$$Y_t - Y_t^\star = XX^\mathsf{T}e_{t-1}, \qquad \mathbb{E}\,\|Y_t - Y_t^\star\|^2 \le C\mathbb{E}\,\|e_{t-1}\|^2.$$

Table 5: Population matrices and secant exponents in the regression applications. Scores and covariance expressions are evaluated at the corresponding population target.

Model	A	Q	q	Limit covariance
Least squares	$\mathbb{E}(XX^\mathsf{T})$	$\mathbb{E}(\varepsilon^2 XX^\mathsf{T})$	0	$\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$
Logistic	$\mathbb{E}\{p_\star(1-p_\star)XX^\mathsf{T}\}$	$\mathbb{E}\{(Y-p_\star)^2XX^\mathsf{T}\}$	0	$\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$
Huber	$\mathbb{E}[XX^\mathsf{T}\mathbf{1}\{\lvert Y - X^\mathsf{T}\theta_c\rvert < c\}]$	$\mathbb{E}\{\psi_c^2 XX^\mathsf{T}\}$	0	$\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$
Quantile, shared endpoints	$\mathbb{E}\{f_{Y\mid X}(X^\mathsf{T}\theta_\tau \mid X)XX^\mathsf{T}\}$	$\mathbb{E}[XX^\mathsf{T}\{\mathbf{1}(Y \le X^\mathsf{T}\theta_\tau) - \tau\}^2]$	1 (upper bound)	$\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$

The average tends to zero. The target $(2+\epsilon_0)$ moment verifies (41), and the remaining second moment conditions follow from the same bounds. Theorem 3.6 proves (56). Under conditional homoskedasticity, $\mathrm{Q_{LS}} = \mathbb{E}\{\mathbb{E}(\varepsilon^2 \mid X)XX^\mathsf{T}\} = \sigma^2 \mathrm{A_{LS}}$. □

Logistic regression

Proof of Corollary C.11. The logistic population risk is

$$R_{\log}(\theta) = \mathbb{E}\{\log(1 + e^{X^\mathsf{T}\theta}) - YX^\mathsf{T}\theta\}.$$

Its gradient is $g_{\log}(\theta) = \mathbb{E}G_{\log}(\theta; X, Y)$ and its Hessian is $\mathrm{H_{log}}(\theta)$ in (57). Because $p'(s) = p(s)\{1-p(s)\}$ is bounded by $1/4$, the risk is smooth. The assumed lower eigenvalue bound makes it strongly convex on $\mathcal{U}$.

Let $w(s) = p(s)\{1-p(s)\}$. Its derivative is bounded. Therefore, for $\theta_1, \theta_2 \in \mathcal{U}$,

$$\begin{aligned}\left\|\mathrm{H_{log}}(\theta_1) - \mathrm{H_{log}}(\theta_2)\right\|_\mathrm{F} &\le C\mathbb{E}\{\|X\|^2 \left|X^\mathsf{T}(\theta_1 - \theta_2)\right|\} \\ &\le C\mathbb{E}\|X\|^3 \|\theta_1 - \theta_2\|.\end{aligned}$$

Thus the population Jacobian is Lipschitz. It follows from the integral identity

$$g_{\log}(\theta_{\log} + e) - \mathrm{A_{log}}e = \int_0^1 \{\mathrm{H_{log}}(\theta_{\log} + se) - \mathrm{A_{log}}\}e\, ds$$

that the mean remainder is $O(\|e\|^2)$.

The sample score is pathwise Lipschitz:

$$\|G_{\log}(\theta_1; X, Y) - G_{\log}(\theta_2; X, Y)\| \le \frac{1}{4}\|X\|^2 \|\theta_1 - \theta_2\|.$$

The fourth moment of X therefore verifies Proposition C.6 and also gives $\mathbb{E}\|\zeta(\theta) - \zeta(\theta_{\log})\|^2 \le C\|\theta - \theta_{\log}\|^2$. The bound $\|G_{\log}(\theta; X, Y)\| \le \|X\|$ holds globally. The assumed mean-square rate supplies (T1), and the global quadratic remainder gives (T3) since $n^{-1/2}\sum_{t\le n}\eta_t \to 0$. No confinement is used.

The conditional secant mean is bounded uniformly over all centers by $C\mathbb{E}\|X\|^2$, because the sample-score Lipschitz bound gives $\left\|\mathrm{sym}(D_{tj}V_t^\mathsf{T})\right\|_\mathrm{F} \le C\|X\|_{tj}^2 \|V\|_t^2$. The iterate rate implies $\theta_t \xrightarrow{p} \theta_{\log}$. Continuity of the Jacobian near this target and $\delta_t \to 0$ therefore imply convergence of the conditional mean to $\mathrm{A_{log}}$ in probability; the deterministic bound upgrades this to unconditional L^2 convergence. Together with the global $q = 0$ second moment bound, this verifies the premises of Theorem 3.3 without a local-trajectory assumption.

At the target, the score has covariance $\mathrm{Q_{log}}$ and a fourth moment. The ordinary iid central limit theorem verifies the target martingale conditions. The secant has $q = 0$, so Theorems 3.2 and 3.3 give (58)–(59). The pathwise Lipschitz bound and the mean-square rate verify (40); the target fourth moment is more than sufficient for (41). Theorem 3.6 proves (60).

Under correct Bernoulli specification,

$$\mathbb{E}\{(Y - p_{\theta_{\log}}(X))^2 \mid X\} = p_{\theta_{\log}}(X)\{1 - p_{\theta_{\log}}(X)\},$$

which gives $\mathrm{Q_{log}} = \mathrm{A_{log}}$. □

Huber regression

Proof of Corollary C.12. Let ρ_c be the Huber loss with derivative ψ_c. Use the anchored population risk

$$\widetilde{R}_c(\theta) = \mathbb{E}\{\rho_c(Y - X^\mathsf{T}\theta) - \rho_c(Y - X^\mathsf{T}\theta_c)\}.$$

It is finite because ρ_c is Lipschitz and $\mathbb{E}\|X\| < \infty$. Since ψ_c is odd, its gradient is $g_c(\theta) = \mathbb{E}\{X\psi_c(X^\mathsf{T}\theta - Y)\}$. For fixed X, the conditional density assumption allows differentiation under

the integral and gives $\frac{d}{ds}\mathbb{E}\{\psi_c(s-Y) \mid X\} = \mathbb{P}(|Y-s| < c \mid X)$. Consequently, $Dg_c(\theta) = \mathrm{H}_c(\theta)$ in (61). The density envelope implies, for $\theta_1, \theta_2 \in \mathcal{U}$,

$$\begin{aligned}&\left|\mathbb{P}(|Y - X^\mathsf{T}\theta_1| < c \mid X) - \mathbb{P}(|Y - X^\mathsf{T}\theta_2| < c \mid X)\right| \\ &\qquad \le 2\bar{f}_c(X)\left|X^\mathsf{T}(\theta_1 - \theta_2)\right|.\end{aligned}$$

It follows that

$$\|\mathrm{H}_c(\theta_1) - \mathrm{H}_c(\theta_2)\|_{\mathrm{F}} \le 2\mathbb{E}\{\bar{f}_c(X)\|X\|^3\}\|\theta_1 - \theta_2\|.$$

Thus the Jacobian is Lipschitz, the anchored risk is smooth and strongly convex on $\mathcal{U}$, and the population-score remainder is $O(\|e\|^2)$.

The function ψ_c is one-Lipschitz, so $\|G_c(\theta_1; X, Y) - G_c(\theta_2; X, Y)\| \le \|X\|^2\|\theta_1 - \theta_2\|$. These global bounds and the fourth moment of X give $q = 0$, quadratic-mean score replacement, and uniformly bounded score second moments. The assumed mean-square rate supplies (T1). Choose $a > 0$ such that the ball of radius a about θ_c lies in $\mathcal{U}$. With $r(e) = g_c(\theta_c + e) - \mathrm{A}_c e$, the local expansion gives $\|r(e)\| \le C\|e\|^2$ on that ball. Outside it, boundedness of g_c gives $\|r(e)\| \le C(1 + \|e\|) \le C_a\|e\|^2$. Thus $\mathbb{E}\|r(e_t)\| = O(\eta_t)$ globally, proving (T3).

The conditional secant mean is bounded by $C\mathbb{E}\|X\|^2$ for every center. Since $\theta_t \xrightarrow{p} \theta_c$ and the perturbations shrink, local continuity of H_c gives convergence of that mean to A_c in probability and, by boundedness, in L^2. The global secant second moment bound then verifies Theorem 3.3. This argument neither assumes confinement nor infers it from local curvature.

At the target, the score covariance is Q_c, and its fourth moment is finite. Applying Theorems 3.2 and 3.3 proves (62)–(63). The same Lipschitz and moment bounds verify the conditions of Theorem 3.6, which gives (64). □

C.5.4 Quantile regression verification

Population expansion and local score continuity

Lemma C.37 (Quantile mean expansion)**.** *Under Assumption 3.4, for e such that $\theta_\tau + e \in \mathcal{U}$,*

$$g_\tau(\theta_\tau + e) = \mathrm{A}_\tau e + r_\tau(e), \qquad \|r_\tau(e)\| \le C\|e\|^{1+\nu}. \tag{105}$$

Proof. Let $q_\star = X^\mathsf{T}\theta_\tau$ and $a = X^\mathsf{T}e$. Since $g_\tau(\theta_\tau) = 0$ by (30),

$$\begin{aligned}g_\tau(\theta_\tau + e) &= g_\tau(\theta_\tau + e) - g_\tau(\theta_\tau) \\ &= \mathbb{E}\left[X\{F_{Y|X}(q_\star + a \mid X) - F_{Y|X}(q_\star \mid X)\}\right].\end{aligned}$$

For every realization of X, the fundamental theorem of calculus gives

$$F_{Y|X}(q_\star + a \mid X) - F_{Y|X}(q_\star \mid X) = f_{Y|X}(q_\star \mid X)a + R_X(a),$$

where

$$\begin{aligned}|R_X(a)| &= \left|\int_0^a \{f_{Y|X}(q_\star + s \mid X) - f_{Y|X}(q_\star \mid X)\}\,ds\right| \\ &\le L_f(X)\int_0^{|a|} s^\nu ds \le CL_f(X)\,|a|^{1+\nu}.\end{aligned}$$

The leading term has expectation $\mathrm{A}_\tau e$. The remainder satisfies

$$\begin{aligned}\left\|\mathbb{E}\{XR_X(X^\mathsf{T}e)\}\right\| &\le C\mathbb{E}\{L_f(X)\|X\|\left|X^\mathsf{T}e\right|^{1+\nu}\} \\ &\le C\mathbb{E}\{L_f(X)\|X\|^{2+\nu}\}\|e\|^{1+\nu},\end{aligned}$$

which is finite by (34). This proves (105). □

Lemma C.38 (Local quadratic-mean continuity of the direct quantile score)**.** *Under Assumption 3.4, for any θ_1, θ_2 in the parameter tube,*

$$\mathbb{E}\|G_\tau(\theta_1; X, Y) - G_\tau(\theta_2; X, Y)\|^2 \le C\|\theta_1 - \theta_2\|. \tag{106}$$

The same order holds for the corresponding centered score difference.

Proof. Put $e = \theta_1 - \theta_2$. The two indicators differ only when Y lies between $X^\mathsf{T}\theta_1$ and $X^\mathsf{T}\theta_2$. Conditional on X, the density envelope bounds the probability of this event by $\bar{f}(X)\left|X^\mathsf{T}e\right| \le \bar{f}(X)\,\|X\|\,\|e\|$. On the event, the squared score difference equals $\|X\|^2$. Therefore

$$\mathbb{E}\,\|G_\tau(\theta_1; X, Y) - G_\tau(\theta_2; X, Y)\|^2 \le \mathbb{E}\{\bar{f}(X)\,\|X\|^3\}\,\|e\|\,,$$

which proves (106). If Z is any square-integrable random vector, then $\mathbb{E}\,\|Z - \mathbb{E}Z\|^2 \le \mathbb{E}\,\|Z\|^2$. Apply this inequality to the raw score difference to obtain the centered version. □

Quantile score secant moments

Proof of Proposition C.13. Fix $\theta = \theta_{t-1}$ and write $q = X^\mathsf{T}\theta, \quad a = X^\mathsf{T}V_t$. For one auxiliary observation, the $-\tau X$ terms cancel, so conditional on $(X, V_t, \mathcal{F}_{t-1})$,

$$\mathbb{E}(D_{t1} \mid X, V_t, \mathcal{F}_{t-1}) = X\frac{F_{Y|X}(q + \delta_t a \mid X) - F_{Y|X}(q - \delta_t a \mid X)}{2\delta_t}.$$

By a change of variables,

$$\frac{F(q + \delta_t a \mid X) - F(q - \delta_t a \mid X)}{2\delta_t} = \frac{a}{2}\int_{-1}^{1} f(q + s\delta_t a \mid X)\,ds. \tag{107}$$

Split the density in (107) into $f(q \mid X)$ plus a remainder. The leading part of the one-observation matrix is

$$\mathrm{sym}\{X(X^\mathsf{T}V_t)V_t^\mathsf{T}\}f(q \mid X) = \mathrm{sym}\{XX^\mathsf{T}V_tV_t^\mathsf{T}\}f(q \mid X).$$

Conditional on X, isotropy and symmetry of XX^T give

$$\begin{aligned}\mathbb{E}_{V_t}\,\mathrm{sym}(XX^\mathsf{T}V_tV_t^\mathsf{T}) &= \frac{1}{2}\{XX^\mathsf{T}\mathbb{E}(V_tV_t^\mathsf{T}) + \mathbb{E}(V_tV_t^\mathsf{T})XX^\mathsf{T}\}\\ &= XX^\mathsf{T}.\end{aligned}$$

Thus the leading conditional mean after averaging over (X, V_t) is $\mathrm{H}_\tau(\theta) = \mathbb{E}\{f_{Y|X}(X^\mathsf{T}\theta \mid X)XX^\mathsf{T}\}$.

For the finite-secant remainder, (33) yields $|f(q + s\delta_t a \mid X) - f(q \mid X)| \le L_f(X)\delta_t^\nu\,|a|^\nu$. After multiplying by XaV_t^T and symmetrizing, its Frobenius norm is bounded by

$$C\delta_t^\nu L_f(X)\,\|X\|\,|a|^{1+\nu}\,\|V_t\| \le C\delta_t^\nu L_f(X)\,\|X\|^{2+\nu}\,\|V_t\|^{2+\nu}\,.$$

The expectation is finite, so the finite-secant mean error is $O(\delta_t^\nu)$. Moreover,

$$\begin{aligned}\|\mathrm{H}_\tau(\theta) - \mathrm{A}_\tau\|_\mathrm{F} &\le \mathbb{E}\left[L_f(X)\left|X^\mathsf{T}(\theta - \theta_\tau)\right|^\nu\|X\|^2\right]\\ &\le C\,\|\theta - \theta_\tau\|^\nu\,.\end{aligned}$$

Combining the last two bounds proves (65).

We next prove the variance bound. Let

$$\mathcal{C}_t = \{Y \text{ lies between } q - \delta_t a \text{ and } q + \delta_t a\}.$$

The score difference is zero outside $\mathcal{C}_t$. On $\mathcal{C}_t$,

$$\left\|\mathrm{sym}(D_{t1}V_t^\mathsf{T})\right\|_\mathrm{F}^2 \le \frac{C}{\delta_t^2}\,\|X\|^2\,\|V_t\|^2\,.$$

Conditional on (X, V_t),

$$\begin{aligned}\mathbb{P}(\mathcal{C}_t \mid X, V_t) &\le 2\delta_t\bar{f}(X)\left|X^\mathsf{T}V_t\right|\\ &\le 2\delta_t\bar{f}(X)\,\|X\|\,\|V_t\|\,.\end{aligned}$$

It follows that

$$\mathbb{E}\left\|\mathrm{sym}(D_{t1}V_t^\mathsf{T})\right\|_\mathrm{F}^2 \le \frac{C}{\delta_t}\mathbb{E}\{\bar{f}(X)\,\|X\|^3\,\|V_t\|^3\} \le \frac{C}{\delta_t}.$$

Conditional averaging of R_t independent observations reduces the sampling variance to $C/(R_t\delta_t)$. The variance of the conditional mean with respect to the bounded direction is $O(1)$ by (34). The law of total variance proves (66). □

Independent endpoint samples

Proof of Proposition C.14. Because $\|G_\tau(\vartheta; X, Y)\| \le \|X\|$ for every ϑ, boundedness of V gives, uniformly over the parameter tube,

$$\mathbb{E}\left\|\mathrm{K}_\delta^{\mathrm{ind}}\right\|_{\mathrm{F}}^2 \le \frac{C}{\delta^2}\mathbb{E}\{\|X\|^2\|V\|^2\} \le C\delta^{-2},$$

which proves (68). Set

$$Z_\delta = G_\tau(\theta+\delta V; Z^+) - G_\tau(\theta-\delta V; Z^-), \qquad Z_0 = G_\tau(\theta; Z^+) - G_\tau(\theta; Z^-).$$

By Lemma C.38, boundedness of V, and conditioning on V,

$$\begin{aligned}\mathbb{E}\|Z_\delta - Z_0\|^2 &\le 2\mathbb{E}\left\|G_\tau(\theta+\delta V; Z^+) - G_\tau(\theta; Z^+)\right\|^2 \\ &\quad + 2\mathbb{E}\left\|G_\tau(\theta-\delta V; Z^-) - G_\tau(\theta; Z^-)\right\|^2 \longrightarrow 0.\end{aligned}$$

Thus $Z_\delta \to Z_0$ in L^2. The limit Z_0 is independent of V, has mean zero, and satisfies

$$\mathbb{E}\|Z_0\|^2 = 2\,\mathrm{tr}\{\mathrm{Var}(G_\tau(\theta; Z))\}. \tag{108}$$

For deterministic vectors z, v,

$$\left\|\mathrm{sym}(zv^\mathsf{T})\right\|_{\mathrm{F}}^2 = \frac{1}{2}\{\|z\|^2\|v\|^2 + (z^\mathsf{T}v)^2\}. \tag{109}$$

Since $\mathbb{E}(VV^\mathsf{T}) = \mathrm{I}_d$, $\mathbb{E}\|V\|^2 = d$, and Z_0 is independent of V, equations (108)–(109) give

$$\begin{aligned}\mathbb{E}\left\|\mathrm{sym}(Z_0V^\mathsf{T})\right\|_{\mathrm{F}}^2 &= \frac{d+1}{2}\mathbb{E}\|Z_0\|^2 \\ &= (d+1)\,\mathrm{tr}\{\mathrm{Var}(G_\tau(\theta; Z))\}.\end{aligned}$$

Boundedness of V and L^2 convergence of Z_δ imply $\mathrm{sym}(Z_\delta V^\mathsf{T}) \longrightarrow \mathrm{sym}(Z_0V^\mathsf{T})$ in L^2. By (67), $\delta\mathrm{K}_\delta^{\mathrm{ind}} = \frac{1}{2}\mathrm{sym}(Z_\delta V^\mathsf{T})$. Moreover $\delta\mathbb{E}\mathrm{K}_\delta^{\mathrm{ind}} \to 0$, because $\mathbb{E}\{\mathrm{sym}(Z_0V^\mathsf{T})\} = 0$ and the displayed L^2 convergence also holds in L^1. Therefore

$$\delta^2\mathbb{E}\left\|\mathrm{K}_\delta^{\mathrm{ind}} - \mathbb{E}\mathrm{K}_\delta^{\mathrm{ind}}\right\|_{\mathrm{F}}^2 \longrightarrow \frac{1}{4}(d+1)\,\mathrm{tr}\{\mathrm{Var}(G_\tau(\theta; Z))\},$$

which is (69). Independence across endpoint pairs divides this sampling variance by R_t. □

Direct-score quantile theorem

Proof of Theorem 3.5. We verify the transfer conditions. Condition (Q5) supplies the deterministic almost-sure localization required by the local alternative in Proposition C.3. By Lemma C.37, the population score has derivative A_τ and remainder $O(\|e\|^{1+\nu})$. The mean-square iterate rate and $\alpha(1+\nu) > 1$ imply (15) through that proposition.

The direct score is unbiased, so $b_t \equiv 0$. By Lemma C.38, the centered local score difference has second moment $O(\|e\|)$; hence (16) holds. At the target, $\zeta_t^\star = G_\tau(\theta_\tau; X_t, Y_t)$, because (30) gives mean zero. Its covariance is the matrix S_τ in (37). The $(2+\epsilon)$ moment in (34) gives the Lindeberg condition, and iid sampling gives the target quadratic-variation limit.

By Proposition C.13, the shared observation curvature estimate has $q = 1$. The mean-square rate and $\nu \le 1$ imply

$$\mathbb{E}\left\|\mathbb{E}(\widehat{\mathrm{H}}_t \mid \mathcal{F}_{t-1}) - \mathrm{A}_\tau\right\|_{\mathrm{F}}^2 \le C\{\mathbb{E}\|e_{t-1}\|^{2\nu} + \delta_t^{2\nu}\} \longrightarrow 0.$$

Thus (65) verifies (24). The condition $(\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha$ and Theorem 3.3 therefore give dynamic stabilization, $\mathrm{B}_t \xrightarrow{p} \mathrm{A}_\tau$, and $\mathrm{P}_t \xrightarrow{p} \mathrm{A}_\tau^{-1}$. Theorem 3.2 proves (36).

For the sandwich estimator, Lemma C.38 and the mean-square rate give

$$\frac{1}{n}\sum_{t=1}^n \mathbb{E}\|Y_t - Y_t^\star\|^2 \le \frac{C}{n}\sum_{t=1}^n \mathbb{E}\|e_{t-1}\| \longrightarrow 0.$$

The target has a uniform $(2+\epsilon)$ moment, so (41) follows from the sufficient condition in Theorem 3.6. The average second moment bound and (42) are immediate. Thus Theorem 3.6 gives (38).

Under (31), conditional on X the indicator in the target score is Bernoulli with success probability τ. Therefore $\mathbb{E}[\{\mathbf{1}(Y \le X^{\mathsf{T}}\theta_\tau) - \tau\}^2 \mid X] = \tau(1-\tau)$, which proves (39). For independent endpoints, Proposition C.14 gives $q = 2$, leading to $(2\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha$. □

C.5.5 Zeroth-order verification

Proof of the smoothing-bias phase

Proof of Proposition C.17. The assumed integrability makes the finite prefix before the trajectory enters $\mathcal{U}$ contribute $O(n^{-1/2})$. For every remaining t, the expansion (72) gives

$$\begin{aligned}&\|b_{h_t}(\theta_{t-1}) - b_{h_t}(\theta_\star)\| \\ &\quad\le h_t^p \|B(\theta_{t-1}) - B(\theta_\star)\| + \|r_{h_t}(\theta_{t-1})\| + \|r_{h_t}(\theta_\star)\| \\ &\quad\le C h_t^p \|\theta_{t-1} - \theta_\star\|^\xi + 2h_t^p \omega(h_t).\end{aligned}$$

Because $\xi \le 2$, Lyapunov's inequality and (13) imply

$$\mathbb{E}\|\theta_{t-1} - \theta_\star\|^\xi \le \{\mathbb{E}\|\theta_{t-1} - \theta_\star\|^2\}^{\xi/2} \le C\eta_{t-1}^{\xi/2}.$$

Take expectations, sum over t, and divide by $\sqrt{n}$ to obtain (73).

It remains to verify the stated polynomial-rate consequence. Put $a = p\kappa$. The first sum in (73) is bounded, up to a constant, by $\frac{1}{\sqrt{n}}\sum_{t=1}^n t^{-(a+\alpha\xi/2)}$. If $a \ge 1/2$, then $a + \alpha\xi/2 > 1/2$; the integral comparison for power sums shows that the display tends to zero.

For the second sum, first suppose $a > 1/2$. Since $\omega(h_t) \to 0$, it is eventually bounded, and $\frac{1}{\sqrt{n}}\sum_{t=1}^n t^{-a}\omega(h_t) \longrightarrow 0$. If $a = 1/2$, fix $\varepsilon > 0$ and choose T such that $|\omega(h_t)| \le \varepsilon$ for $t \ge T$. The finite prefix divided by $\sqrt{n}$ vanishes, whereas $\limsup_{n\to\infty} \frac{1}{\sqrt{n}}\sum_{t=T}^n t^{-1/2}|\omega(h_t)| \le 2\varepsilon$. Letting $\varepsilon \downarrow 0$ proves convergence. Thus both terms vanish whenever $p\kappa \ge 1/2$, and (17) follows. □

Proof of Theorem C.18. The proof of Theorem 3.2 before imposing a finite limit on the target bias gives

$$\sqrt{n}(\bar\theta_n - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^n \zeta_t^\star - \mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^n b_t(\theta_\star) + o_p(1). \tag{110}$$

Put $a = p\kappa$. By (71),

$$\frac{1}{\sqrt{n}}\sum_{t=1}^n b_t(\theta_\star) = \frac{h_0^p B}{\sqrt{n}}\sum_{t=1}^n t^{-a} + \frac{1}{\sqrt{n}}\sum_{t=1}^n \rho_t^b.$$

We first control the remainder without an additional uniformity condition. Given $\varepsilon > 0$, choose T so that $\mathbb{E}\left\|\rho_t^b\right\| \le \varepsilon h_0^p t^{-a}$ for $t \ge T$. Then

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^n \rho_t^b\right\| \le \frac{1}{\sqrt{n}}\sum_{t<T}\mathbb{E}\left\|\rho_t^b\right\| + \frac{\varepsilon h_0^p}{\sqrt{n}}\sum_{t=T}^n t^{-a}. \tag{111}$$

The finite-prefix term vanishes. Equation (111) implies that the remainder is $o_p(1)$ when $a \ge 1/2$ and is $o_p(n^{1/2-a})$ when $a < 1/2$.

Before treating the subcritical case, note that (T1) gives

$$\mathbb{E}\left\|n^a(\bar\theta_n - \theta_\star)\right\| \le \frac{n^a}{n}\sum_{t=1}^n \{\mathbb{E}\|e_{t-1}\|^2\}^{1/2} = O(n^{a-\alpha/2}).$$

If $a < \alpha/2$, the left side tends to zero. This contradicts the nonzero limit implied below when $B \ne 0$, so, under the other transfer conditions used by that representation, (T1) and the target-bias expansion cannot both hold in that range.

If $a > 1/2$, then $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} t^{-a} \longrightarrow 0$. If $a = 1/2$, the integral comparison for the harmonic square-root sum gives $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} t^{-1/2} \longrightarrow 2$. If $a < 1/2$, $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} t^{-a} \sim \frac{n^{1/2-a}}{1-a}$. The target martingale term in (110) converges to $N(0, \mathrm{A}^{-1}\mathrm{Q}_m\mathrm{A}^{-\mathsf{T}})$ and is therefore tight. Substituting the three deterministic asymptotics into (110) proves parts (i) and (ii). In part (iii), divide (110) by $n^{1/2-a}$. The martingale term is $O_p(1)$ and the transfer remainder is $o_p(1)$, so both vanish on this larger scale, while (111) gives $n^{a-1/2} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \rho_t^b = o_p(1)$. Consequently,

$$n^a(\bar{\theta}_n - \theta_\star) \xrightarrow{p} -\frac{h_0^p}{1-a} \mathrm{A}^{-1} B \neq 0.$$

The root-n sequence therefore diverges in the direction $\mathrm{A}^{-1}B$ and is not tight. □

Symmetric-difference bias order

Proof of Proposition C.20. It is enough to consider one direction U. Taking expectation over Z first and applying the fundamental theorem of calculus along $s \mapsto \theta_\star + shU$ gives

$$\mathbb{E}_Z \left[\frac{L(\theta_\star + hU; Z) - L(\theta_\star - hU; Z)}{2h} U \right] = \frac{1}{2} \int_{-1}^{1} UU^{\mathsf{T}} g(\theta_\star + shU)\, ds. \tag{112}$$

The remainder $r(v) = g(\theta_\star + v) - \mathrm{A}v$ from (12) satisfies $\|r(v)\| \le \|v\|\, \omega_g(\|v\|)$. The linear term vanishes because $(h/2) \int_{-1}^{1} s\, UU^{\mathsf{T}} \mathrm{A} U\, ds = 0$. For the remainder, boundedness of U and monotone replacement of ω_g by its nondecreasing envelope give

$$\left\| \frac{1}{2} \int_{-1}^{1} UU^{\mathsf{T}} r(shU)\, ds \right\| \le \frac{1}{2} \int_{-1}^{1} \|U\|^2\, |s|\, h\, \|U\|\, \omega_g(|s|\, h\, \|U\|)\, ds \le Ch\omega_g(\bar{u}h).$$

Expectation over U and averaging a fixed number of directions prove (74). Summing this bound gives (75). If $\omega_g(a) \le Ca^\nu$, the same display is $O(h^{1+\nu})$, which proves (76). The unbounded-direction statement follows by the same calculation under the stated domination. □

Explicit second-order bias

Proof of Proposition C.21. Starting from the line-integral identity (112), insert (77). The linear term integrates to zero because $\int_{-1}^{1} s\, ds = 0$. The quadratic term equals

$$\frac{1}{2} \int_{-1}^{1} UU^{\mathsf{T}} \frac{1}{2} \mathcal{T}[shU, shU] ds = \frac{h^2}{4} UU^{\mathsf{T}} \mathcal{T}[U, U] \int_{-1}^{1} s^2 ds = \frac{h^2}{6} UU^{\mathsf{T}} \mathcal{T}[U, U].$$

The remainder is $o(h^2)$ uniformly over the bounded support of U, and dominated convergence permits expectation over U. This proves (78).

For quantile regression, put $a = X^{\mathsf{T}} v$. Integrating (79) from 0 to a gives

$$F_{Y|X}(q_\star + a \mid X) - F_{Y|X}(q_\star \mid X) = f_{Y|X}(q_\star \mid X) a + \frac{1}{2} \dot{f}_\star(X) a^2 + R_X(a),$$
$$|R_X(a)| \le C L_2(X)\, |a|^{2+\gamma}.$$

Multiplication by X and expectation yield

$$g_\tau(\theta_\tau + v) = \mathrm{A}_\tau v + \frac{1}{2} \mathbb{E}\{\dot{f}_\star(X) X (X^{\mathsf{T}} v)^2\} + r_2(v),$$

where

$$\|r_2(v)\| \le C \mathbb{E}\{L_2(X)\, \|X\|^{3+\gamma}\}\, \|v\|^{2+\gamma} = o(\|v\|^2).$$

This identifies (81); substitution into (78) gives (82).

Under the additional uniform assumptions, repeat the preceding calculation around an arbitrary θ in the stated neighborhood. With $q_\theta = X^\mathsf{T}\theta$, integration of the uniform density expansion gives, uniformly over bounded v,

$$g_\tau(\theta + v) = g_\tau(\theta) + \mathrm{H}_\tau(\theta)v + \frac{1}{2}\mathcal{T}_\tau(\theta)[v, v] + R_\theta(v), \tag{113}$$

$$\mathcal{T}_\tau(\theta)[v, v] = \mathbb{E}\{\dot{f}_\theta(X)X(X^\mathsf{T}v)^2\}, \qquad \sup_\theta \|R_\theta(v)\| \le C\|v\|^{2+\gamma}. \tag{114}$$

The remainder bound follows from the common envelope $L_2(X)$ and $\mathbb{E}\{L_2(X)\|X\|^{3+\gamma}\} < \infty$.

Use the line-integral representation (112) with $\theta_\star$ replaced by θ. The constant term becomes $\mathbb{E}_U(UU^\mathsf{T})g_\tau(\theta) = g_\tau(\theta)$, the term linear in s integrates to zero, and the quadratic term equals $h^2 B_\tau(\theta)$ with $B_\tau(\theta)$ defined in (84). Equations (113)–(114) and boundedness of U give the uniform remainder bound $\sup_\theta \left\|b_h(\theta) - h^2 B_\tau(\theta)\right\| \le Ch^{2+\gamma}$. Thus (72) holds with $p = 2$ and $\omega(h) = Ch^\gamma$.

Finally, (83) and boundedness of U imply

$$\|B_\tau(\theta_1) - B_\tau(\theta_2)\| \le C\mathbb{E}\{L_{\dot{f}}(X)\|X\|^{3+\xi}\}\|\theta_1 - \theta_2\|^\xi,$$

which is finite by assumption. This proves the uniform local-bias assertions. □

Check-loss finite difference approximation Use the anchored risk

$$R_\tau(\theta) = \mathbb{E}\{\rho_\tau(Y - X^\mathsf{T}\theta) - \rho_\tau(Y - X^\mathsf{T}\theta_\tau)\}.$$

The check loss is one-Lipschitz, so the integrand is bounded in absolute value by $\|X\|\,\|\theta - \theta_\tau\|$. Thus (Q2) makes this expectation finite without a response moment. On the density-controlled tube, dominated differentiation gives $\nabla R_\tau = g_\tau$ and the Jacobian in (32): the first derivative is dominated by $\|X\|$, and the second by $\bar{f}(X)\|X\|^2$. The anchor cancels in every paired difference. Consequently Propositions C.20 and C.21 apply to this risk even if $\mathbb{E}\rho_\tau(Y - X^\mathsf{T}\theta_\tau) = \infty$.

Lemma C.39 (Check-loss L^2 approximation)**.** *Let $L(\theta; X, Y) = \rho_\tau(Y - X^\mathsf{T}\theta)$ and, for a bounded isotropic direction U, define*

$$D_h(\theta; X, Y, U) = \frac{L(\theta + hU; X, Y) - L(\theta - hU; X, Y)}{2h}U.$$

Under Assumption 3.4, uniformly for θ in a sufficiently small neighborhood of θ_τ,

$$\mathbb{E}\left\|D_h(\theta) - UU^\mathsf{T}G_\tau(\theta; X, Y)\right\|^2 \le Ch, \tag{115}$$

$$\mathbb{E}\left\|\{D_h(\theta) - \mathbb{E}D_h(\theta)\} - \{D_h(\theta_\tau) - \mathbb{E}D_h(\theta_\tau)\}\right\|^2 \le C\{\|\theta - \theta_\tau\| + h\}. \tag{116}$$

Proof. Write $q_\theta = Y - X^\mathsf{T}\theta$, $a = X^\mathsf{T}U$. The check loss is linear on each side of zero. If $|q_\theta| > h|a|$, the entire interval between $q_\theta - ha$ and $q_\theta + ha$ lies on one linear branch. On this event, direct calculation gives

$$D_h(\theta; X, Y, U) = UU^\mathsf{T}G_\tau(\theta; X, Y). \tag{117}$$

Both sides of (117) have norm at most $C\|X\|\,\|U\|^2$. Conditional on (X, U),

$$\mathbb{P}(|q_\theta| \le h|a| \mid X, U) \le 2h\bar{f}(X)|a| \le 2h\bar{f}(X)\|X\|\,\|U\|.$$

Consequently,

$$\mathbb{E}\left\|D_h(\theta) - UU^\mathsf{T}G_\tau(\theta)\right\|^2 \le Ch\mathbb{E}\{\bar{f}(X)\|X\|^3\|U\|^5\} \le Ch,$$

which proves (115).

For the local difference, insert and subtract the direct score approximations at θ and θ_τ:

$$\begin{aligned}\mathbb{E}\|D_h(\theta) - D_h(\theta_\tau)\|^2 \le\ & 3\mathbb{E}\left\|D_h(\theta) - UU^\mathsf{T}G_\tau(\theta)\right\|^2 \\ & + 3\mathbb{E}\left\|UU^\mathsf{T}\{G_\tau(\theta) - G_\tau(\theta_\tau)\}\right\|^2\end{aligned}$$

$$+ 3\mathbb{E}\left\|UU^{\mathsf{T}}G_\tau(\theta_\tau) - D_h(\theta_\tau)\right\|^2.$$

The first and third terms are $O(h)$. Boundedness of U and Lemma C.38 make the middle term $O(\|\theta - \theta_\tau\|)$. Finally, centering cannot increase the second moment by more than a constant factor: for any random vector Z, $\mathbb{E}\|Z - \mathbb{E}Z\|^2 \le \mathbb{E}\|Z\|^2$. Apply this to $Z = D_h(\theta) - D_h(\theta_\tau)$ to prove (116). □

Proof of Corollary C.22. By Lemma C.39, the target finite difference score converges in L^2 to the random-direction transform of the direct quantile score, and its centered local increments satisfy the required quadratic-mean condition. Therefore the target covariance is the established Q_m in Proposition C.16.

The check loss is one-Lipschitz in its scalar residual, so boundedness of U gives

$$\|D_h(\theta; X, Y, U)\| \le \|U\| \frac{2h\left|X^{\mathsf{T}}U\right|}{2h} \le \|X\|\,\|U\|^2.$$

The moment envelope in (34) therefore supplies a uniform $(2+\epsilon')$ moment for some $\epsilon' > 0$, and hence the martingale Lindeberg condition. Proposition C.20 gives the upper order $O(h^{1+\nu})$; the assumed actual expansion identifies p and its leading vector. Under the differentiable-density conditions, Proposition C.21 gives $p = 2$ and $B = B_\tau$ explicitly. The curvature preconditioner is unchanged and satisfies the direct score tuning condition by Proposition C.13 and Theorem 3.3. Applying Theorem C.18 gives the stated three regimes.

For covariance estimation, (116), the mean-square rate, and $h_t \to 0$ imply (40). The bound $\|D_h(\theta)\| \le C\|X\|$ gives the average second moment condition and a uniform $(2+\epsilon')$ target moment. Finally, (115) implies convergence of the conditional target second moment to Q_m. Hence Theorem 3.6 gives $\widehat{\Sigma}_n \xrightarrow{p} \mathrm{A}_\tau^{-1}\mathrm{Q}_m\mathrm{A}_\tau^{-1}$ in the centered and critical regimes. □

C.5.6 Proof of recursive covariance consistency

Proof of Theorem 3.6. Because $\widehat{\mathrm{Q}}_0 = 0$, recursion (9) gives $\widehat{\mathrm{Q}}_n = \frac{1}{n}\sum_{t=1}^n Y_t Y_t^{\mathsf{T}}$. Decompose

$$\widehat{\mathrm{Q}}_n - \mathrm{Q} = \frac{1}{n}\sum_{t=1}^n \{Y_t Y_t^{\mathsf{T}} - Y_t^\star Y_t^{\star\mathsf{T}}\} \tag{118}$$

$$+ \frac{1}{n}\sum_{t=1}^n \left[Y_t^\star Y_t^{\star\mathsf{T}} - \mathbb{E}(Y_t^\star Y_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1})\right] \tag{119}$$

$$+ \left\{\frac{1}{n}\sum_{t=1}^n \mathbb{E}(Y_t^\star Y_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1}) - \mathrm{Q}\right\}. \tag{120}$$

For vectors x, y,

$$\left\|xx^{\mathsf{T}} - yy^{\mathsf{T}}\right\|_{\mathrm{F}} \le (\|x\| + \|y\|)\,\|x - y\|.$$

Hence, by Cauchy-Schwarz first in probability and then over t,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^n \{Y_t Y_t^{\mathsf{T}} - Y_t^\star Y_t^{\star\mathsf{T}}\}\right\|_{\mathrm{F}} \le \left\{\frac{1}{n}\sum_{t=1}^n \mathbb{E}(\|Y_t\| + \|Y_t^\star\|)^2\right\}^{1/2}\left\{\frac{1}{n}\sum_{t=1}^n \mathbb{E}\|Y_t - Y_t^\star\|^2\right\}^{1/2} \longrightarrow 0.$$

Thus (118) is $o_p(1)$. Terms (119) and (120) are $o_p(1)$ by (41) and (42). This proves $\widehat{\mathrm{Q}}_n \xrightarrow{p} \mathrm{Q}$.

We also verify the stated sufficient condition for (41). Let $\delta = \min(\epsilon, 2)$ and $p = 1 + \delta/2 \in (1, 2]$. For each matrix entry,

$$D_{t,jk} = Y_{t,j}^\star Y_{t,k}^\star - \mathbb{E}(Y_{t,j}^\star Y_{t,k}^\star \mid \mathcal{F}_{t-1})$$

is a martingale difference and has a uniformly bounded pth moment because $2p = 2 + \delta \le 2 + \epsilon$. The martingale von Bahr-Esseen inequality gives

$$\mathbb{E}\left|\frac{1}{n}\sum_{t=1}^n D_{t,jk}\right|^p \le Cn^{-p}\sum_{t=1}^n \mathbb{E}\,|D_{t,jk}|^p \le Cn^{1-p} \longrightarrow 0.$$

There are finitely many entries, so (41) follows.

Since $\mathrm{B}_n \xrightarrow{p} \mathrm{A}$ and all eigenvalues of B_n are at least ℓ, inversion is continuous and $\mathrm{B}_n^{-1} \xrightarrow{p} \mathrm{A}^{-1}$. Slutsky's theorem gives

$$\mathrm{B}_n^{-1}\widehat{\mathrm{Q}}_n\mathrm{B}_n^{-1} \xrightarrow{p} \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}.$$

Combining this consistency with the centered central limit theorem proves the Wald coverage statement. □

Proof of Corollary C.24. The first conclusion follows from Theorem 3.2: once (T1)–(T6) and the centered target-bias condition hold, the leading term is $-\mathrm{A}^{-1}n^{-1/2}\sum_{t=1}^{n}\zeta_t^\star$ regardless of the limit of P_t. The increment conclusion of Theorem 3.3 verifies (T2) under the stated threshold, while that theorem also gives $\mathrm{B}_n \xrightarrow{p} \mathrm{A}_{\ell,u}$.

The proof of Theorem 3.6 through (118)–(120) does not use $\mathrm{B}_n \xrightarrow{p} \mathrm{A}$ and therefore still gives $\widehat{\mathrm{Q}}_n \xrightarrow{p} \mathrm{Q}$. Continuity of inversion on $\mathcal{S}_{\ell,u}$ and Slutsky's theorem yield

$$\mathrm{B}_n^{-1}\widehat{\mathrm{Q}}_n\mathrm{B}_n^{-1} \xrightarrow{p} \mathrm{A}_{\ell,u}^{-1}\mathrm{Q}\mathrm{A}_{\ell,u}^{-1}.$$

Comparison with the covariance in Theorem 3.2 proves the stated validity condition. If Q is positive definite, equality of the two covariance matrices implies

$$(\mathrm{Q}^{1/2}\mathrm{A}_{\ell,u}^{-1}\mathrm{Q}^{1/2})^2 = (\mathrm{Q}^{1/2}\mathrm{A}^{-1}\mathrm{Q}^{1/2})^2.$$

Both factors are positive definite. Uniqueness of the positive definite square root gives $\mathrm{A}_{\ell,u}^{-1} = \mathrm{A}^{-1}$, hence $\mathrm{A}_{\ell,u} = \mathrm{A}$. The converse is immediate. For singular Q, the displayed covariance equality is the exact condition, but it need not identify the matrices outside the range of Q. □

Proof of Corollary C.26. For a fixed positive definite matrix P satisfying the stability assumptions, the effective inverse drift satisfies $(\mathrm{PA})^{-1}\mathrm{P} = \mathrm{A}^{-1}$. Thus taking $\mathrm{P} = \mathrm{I}_d$ gives ordinary PR averaging, while $\mathrm{P} = \mathrm{A}^{-1}$ gives oracle Newton preconditioning, without changing the leading averaged influence function. A stochastic Newton or conditioned-gradient matrix is covered whenever its effective inverse drift satisfies the moment and increment conditions in Assumption 3.1 and Proposition C.1. Taking $\mathrm{P}_t = \mathrm{I}_d$ and the finite difference primary score gives the averaged Kiefer-Wolfowitz specialization, with covariance Q_m by Proposition C.16. Finally, Proposition C.6 gives $q = 0$ for a pathwise Lipschitz auxiliary score. Then $s_{\mathrm{H}} = (0 \cdot \kappa_{\mathrm{H}} - r)_+ = 0 < 1 - \alpha$, proving the last statement. □

C.5.7 Proof of the conditioning comparison

Proof of Proposition C.28. Part (i) is Theorem 3.2 with zero accumulated target bias. For part (ii), set $z_t = \mathrm{A}^{1/2}e_t$. Then

$$z_t = (\mathrm{I}_d - \eta_t\mathrm{C}_{\mathrm{P}})z_{t-1}, \qquad \mathrm{C}_{\mathrm{P}} = \mathrm{A}^{1/2}\mathrm{P}\mathrm{A}^{1/2}.$$

The matrix C_{P} is symmetric positive definite. Under $0 < \eta_t\lambda_{\max}(\mathrm{C}_{\mathrm{P}}) \le 1$, every eigenvalue of $\mathrm{I}_d - \eta_t\mathrm{C}_{\mathrm{P}}$ lies in $[0, 1)$, so

$$\|z_t\| \le \{1 - \eta_t\lambda_{\min}(\mathrm{C}_{\mathrm{P}})\}\,\|z_{t-1}\| \le \exp\{-\eta_t\lambda_{\min}(\mathrm{C}_{\mathrm{P}})\}\,\|z_{t-1}\|.$$

Iterating gives (85) because $\|z_t\| = \|e_t\|_{\mathrm{A}}$. If $\eta_t = \rho\lambda_{\max}(\mathrm{C}_{\mathrm{P}})^{-1}t^{-\alpha}$, then

$$\lambda_{\min}(\mathrm{C}_{\mathrm{P}})\sum_{s=1}^{t}\eta_s = \frac{\rho}{\kappa(\mathrm{C}_{\mathrm{P}})}\sum_{s=1}^{t}s^{-\alpha}.$$

Finally, $\mathrm{P} = \mathrm{A}^{-1}$ gives $\mathrm{C}_{\mathrm{P}} = \mathrm{I}_d$. □

C.5.8 PROOFS FOR THE FUNCTIONAL LIMIT AND SELF-NORMALIZATION

Lemma C.40 (Uniform quadratic variation from the terminal limit)**.** *For the single time-indexed sequence* $\mathrm{C}_t = \mathbb{E}(\zeta_t^\star \zeta_t^{\star\mathsf{T}} \mid \mathcal{F}_{t-1})$*, the terminal limit* (18) *implies* (86)*.*

Proof. Set $\mathrm{S}_n(r) = n^{-1}\sum_{t\le\lfloor nr\rfloor}\mathrm{C}_t$. For fixed $r > 0$ and $k_n = \lfloor nr\rfloor$, $\mathrm{S}_n(r) = (k_n/n)\{k_n^{-1}\sum_{t\le k_n}\mathrm{C}_t\} \xrightarrow{p} r\mathrm{Q}$; at $r = 0$ this is immediate. For an integer K, let $\epsilon_{n,K} = \max_{0\le j\le K}\|\mathrm{S}_n(j/K) - (j/K)\mathrm{Q}\|_{\mathrm{op}}$. Finite-grid convergence gives $\epsilon_{n,K} \xrightarrow{p} 0$. If $j/K \le r \le (j+1)/K$, positive semidefiniteness of each C_t implies $\mathrm{S}_n(j/K) \preceq \mathrm{S}_n(r) \preceq \mathrm{S}_n((j+1)/K)$. Subtract $r\mathrm{Q}$ and bound the extremal eigenvalues to obtain $\|\mathrm{S}_n(r) - r\mathrm{Q}\|_{\mathrm{op}} \le \epsilon_{n,K} + K^{-1}\|\mathrm{Q}\|_{\mathrm{op}}$. Take the supremum over r, then let $n\to\infty$ and $K\to\infty$. The fixed-sequence premise is essential: a horizon-dependent array with covariance $2\mathrm{Q}$ on its first half and zero on its second half has the terminal limit but not the stated linear variance profile. □

Proof of Proposition C.31. For part (i),

$$\begin{aligned}\frac{1}{\sqrt{n}}\max_{1\le t\le n} a_t\|\mathrm{M}_t e_t\| &\le \eta_0^{-1} n^{(\alpha-1)/2}\max_{1\le t\le n} t^{\alpha/2}\|\mathrm{M}_t e_t\| \\ &= O_p\{n^{(\alpha-1)/2}L_n\} = o_p(1).\end{aligned}$$

For part (ii), the union bound and Markov's inequality give, for every $\varepsilon > 0$,

$$\begin{aligned}&\mathbb{P}\left(\max_{1\le t\le n} a_t\|\mathrm{M}_t e_t\| > \varepsilon\sqrt{n}\right) \\ &\le \frac{C}{n^{p/2}}\sum_{t=1}^n t^{\alpha p}\mathbb{E}\|\mathrm{M}_t e_t\|^p \le Cn^{-p/2}\sum_{t=1}^n t^{\alpha p/2} \\ &\le Cn^{1-p(1-\alpha)/2} \longrightarrow 0,\end{aligned}$$

where the last exponent is negative because $p > 2/(1-\alpha)$. □

Proof of Theorem C.30. Fix $m \le n$ and sum the exact decomposition (99) over $t = 1,\dots,m$, dividing by $\sqrt{n}$. By Lemma C.34, the partial dynamic term is

$$\mathcal{R}_n^{\mathrm{dyn}}(m) = \frac{1}{\sqrt{n}}\left\{a_1\mathrm{M}_1 e_0 - a_m\mathrm{M}_m e_m + \sum_{t=1}^{m-1}(a_{t+1}\mathrm{M}_{t+1} - a_t\mathrm{M}_t)e_t\right\}.$$

Consequently,

$$\begin{aligned}\max_{1\le m\le n}\left\|\mathcal{R}_n^{\mathrm{dyn}}(m)\right\| \le \frac{a_1\|\mathrm{M}_1 e_0\|}{\sqrt{n}} + \frac{1}{\sqrt{n}}\max_{1\le m\le n} a_m\|\mathrm{M}_m e_m\| \\ + \frac{1}{\sqrt{n}}\sum_{t=1}^{n-1}\|(a_{t+1}\mathrm{M}_{t+1} - a_t\mathrm{M}_t)e_t\|.\end{aligned}$$

The first term tends to zero in probability. The second does so by (F3). The expectation of the third tends to zero by the calculation in (96)–(97). Hence the dynamic term is uniformly $o_p(1)$.

Let $\Delta_t = \zeta_t(\theta_{t-1}) - \zeta_t^\star$. It is a martingale difference. Doob's inequality and (T4) yield

$$\mathbb{E}\max_{1\le m\le n}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^m\Delta_t\right\|^2 \le \frac{4}{n}\sum_{t=1}^n\mathbb{E}\|\Delta_t\|^2 \longrightarrow 0.$$

The partial mean-remainder and local-bias terms satisfy

$$\sup_{0\le r\le 1}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor nr\rfloor} r(e_{t-1})\right\| \le \frac{1}{\sqrt{n}}\sum_{t=1}^n\|r(e_{t-1})\|,$$

$$\sup_{0\le r\le 1}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor nr\rfloor}\{b_t(\theta_{t-1})-b_t(\theta_\star)\}\right\|\le\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\|b_t(\theta_{t-1})-b_t(\theta_\star)\|.$$

Their expectations tend to zero by (T3) and (T5). We have therefore proved the uniform representation

$$\sup_{0\le r\le 1}\left\|\mathbb{Z}_n(r)+\mathrm{A}^{-1}\mathbb{U}_n(r)+\mathrm{A}^{-1}\mathbb{B}_n(r)\right\|\xrightarrow{p}0. \tag{121}$$

By Lemma C.40, the conditional Lindeberg condition in (T6), and the martingale functional central limit theorem Hall & Heyde (1980), $\mathbb{U}_n(\cdot)\Longrightarrow \mathrm{Q}^{1/2}W_d(\cdot)$ in $D([0,1],\mathbb{R}^d)$. Condition (F2), (121), and Slutsky's theorem in the product space give (88). The centered statement follows by taking $\mathbb{B}\equiv 0$ and absorbing the minus sign into Brownian motion. □

Proof of Corollary C.32. For each direct score model, $b_t\equiv 0$, so (F2) holds with $\mathbb{B}\equiv 0$. At the target, the primary scores are iid across t and independent of $\mathcal{F}_{t-1}$. Their conditional covariance is the model-specific matrix Q, and hence (86) holds, apart from the deterministic error $|\lfloor nr\rfloor/n-r|\,\|\mathrm{Q}\|_{\mathrm{op}}\le n^{-1}\|\mathrm{Q}\|_{\mathrm{op}}$. The stated $2+\epsilon$ moment conditions imply the conditional Lindeberg condition. The remaining aggregate conditions are verified in the proofs of the corresponding pointwise results, and (F3) is assumed in the corollary. Apply Theorem C.30. □

Lemma C.41 (Positive definiteness of the Brownian normalizer)**.** *For every fixed $\ell\ge 1$,*

$$\mathrm{J}_\ell=\int_0^1\overline{W}_\ell(r)\overline{W}_\ell(r)^{\mathsf{T}}\,dr$$

is positive definite almost surely.

Proof. Choose distinct points $0<r_1<\cdots<r_\ell<1$. The ℓ-vector $(\overline{W}_1(r_1),\dots,\overline{W}_1(r_\ell))$ has a nonsingular Gaussian covariance matrix. The corresponding vectors for the ℓ independent coordinates are iid. Therefore the random matrix with entries $\overline{W}_i(r_j)$ has an absolutely continuous distribution on $\mathbb{R}^{\ell\times\ell}$ and is nonsingular almost surely.

If J_ℓ were singular, there would be a nonzero vector a such that

$$0=a^{\mathsf{T}}\mathrm{J}_\ell a=\int_0^1\{a^{\mathsf{T}}\overline{W}_\ell(r)\}^2\,dr.$$

Continuity would imply $a^{\mathsf{T}}\overline{W}_\ell(r)=0$ for every r, and in particular at $r_1,\dots,r_\ell$. This contradicts nonsingularity of the displayed random matrix. □

Proof of Theorem C.33. The algebraic identity (91) follows from

$$\mathbb{Z}_n(s/n)-\frac{s}{n}\mathbb{Z}_n(1)=\frac{s}{\sqrt{n}}(\bar\theta_s-\bar\theta_n).$$

Since the limit in (89) has continuous paths, convergence in the Skorokhod topology implies the usual continuous-mapping conclusion for the Riemann sum in (91). Hence

$$\mathrm{V}_n\Longrightarrow\Gamma^{1/2}\left\{\int_0^1\overline{W}_d(r)\overline{W}_d(r)^{\mathsf{T}}\,dr\right\}\Gamma^{1/2}. \tag{122}$$

Let L be a nonsingular square root of $\mathrm{R}\Gamma\mathrm{R}^{\mathsf{T}}$. The process $\mathrm{R}\Gamma^{1/2}W_d$ has the same law as $\mathrm{L}W_\ell$. Under the null, the numerator in (92) and the restricted version of (122) therefore converge jointly to

$$\mathrm{L}W_\ell(1)\quad\text{and}\quad\mathrm{L}\mathrm{J}_\ell\mathrm{L}^{\mathsf{T}}.$$

By Lemma C.41, the second matrix is nonsingular almost surely. The continuous mapping theorem gives $W_\ell(1)^{\mathsf{T}}\mathrm{J}_\ell^{-1}W_\ell(1)$, proving (92). The Brownian endpoint is independent of the Brownian bridge. Conditional on the positive definite bridge normalizer, the quadratic-form pivot has a continuous distribution; its unconditional distribution is therefore continuous. Convergence at its critical quantiles proves the confidence-region statement. For a scalar contrast, $c^{\mathsf{T}}\Gamma^{1/2}W_d$ is a scalar Brownian motion multiplied by $(c^{\mathsf{T}}\Gamma c)^{1/2}$; this factor cancels between numerator and denominator. The scalar pivot is continuous and symmetric, which proves (93). □

C.5.9 A PRIMITIVE ITERATE RATE RESULT

The inferential theorem treats the standard mean-square rate as a stability input because global localization is model-specific. The next proposition gives a sufficient route based on descent of the population risk. It does not require Euclidean monotonicity of $\mathrm{P}_t g(\theta)$, which need not follow from positive definiteness of P_t alone.

Proposition C.42. *Let $R : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex, and L-smooth on a convex stability region containing the trajectory and $\theta_\star$, with $g = \nabla R$, $g(\theta_\star) = 0$, and minimizer $\theta_\star$. Suppose $\mathbb{E}\{R(\theta_0) - R(\theta_\star)\} < \infty$, and suppose P_t is $\mathcal{F}_{t-1}$-measurable and, for constants $0 < \underline{p} \le \overline{p} < \infty$, $\underline{p}\mathrm{I}_d \preceq \mathrm{P}_t \preceq \overline{p}\mathrm{I}_d$ almost surely, and the primary oracle satisfies*

$$\mathbb{E}\{\widehat{G}_t(\theta) \mid \mathcal{F}_{t-1}\} = g(\theta) + b_t(\theta),$$

$$\mathbb{E}\left[\left\|\widehat{G}_t(\theta) - \mathbb{E}(\widehat{G}_t(\theta) \mid \mathcal{F}_{t-1})\right\|^2 \mid \mathcal{F}_{t-1}\right] \le C\{1 + \|g(\theta)\|^2\}, \tag{123}$$

$$\|b_t(\theta)\| \le C_b \eta_t^{1/2} \tag{124}$$

throughout the region. For sufficiently small η_0,

$$\mathbb{E}\{R(\theta_t) - R(\theta_\star)\} = O(\eta_t), \qquad \mathbb{E}\,\|\theta_t - \theta_\star\|^2 = O(\eta_t). \tag{125}$$

The proposition applies whenever the displayed update remains in the stated stability region. For a merely local region, proving that the unmodified recursion remains there, or proving the corresponding rate for a projected or truncated modification, is a separate step.

Proof. Write $g_{t-1} = g(\theta_{t-1})$, $b_{t-1} = b_t(\theta_{t-1})$, and $\widehat{G}_t = \widehat{G}_t(\theta_{t-1})$. By L-smoothness,

$$R(\theta_t) \le R(\theta_{t-1}) - \eta_t g_{t-1}^\mathsf{T} \mathrm{P}_t \widehat{G}_t + \frac{L}{2}\eta_t^2 \left\|\mathrm{P}_t \widehat{G}_t\right\|^2.$$

Take conditional expectations. The first-order term becomes

$$\begin{aligned} -\eta_t g_{t-1}^\mathsf{T} \mathrm{P}_t (g_{t-1} + b_{t-1}) &\le -\underline{p}\eta_t \|g_{t-1}\|^2 + \overline{p}\eta_t \|g_{t-1}\| \|b_{t-1}\| \\ &\le -\frac{3\underline{p}}{4}\eta_t \|g_{t-1}\|^2 + C\eta_t \|b_{t-1}\|^2 \\ &\le -\frac{3\underline{p}}{4}\eta_t \|g_{t-1}\|^2 + C\eta_t^2, \end{aligned}$$

where Young's inequality and (124) were used.

For the quadratic term, (123) and (124) imply

$$\begin{aligned} \mathbb{E}(\left\|\mathrm{P}_t \widehat{G}_t\right\|^2 \mid \mathcal{F}_{t-1}) &\le C\{1 + \|g_{t-1}\|^2 + \|b_{t-1}\|^2\} \\ &\le C\{1 + \|g_{t-1}\|^2\}. \end{aligned}$$

For sufficiently small η_t, the resulting $C\eta_t^2 \|g_{t-1}\|^2$ term is absorbed into the negative first-order term. Thus

$$\mathbb{E}\{R(\theta_t) - R(\theta_\star) \mid \mathcal{F}_{t-1}\} \le R(\theta_{t-1}) - R(\theta_\star) - c\eta_t \|g_{t-1}\|^2 + C\eta_t^2. \tag{126}$$

Strong convexity implies the Polyak-Łojasiewicz inequality

$$\|g(\theta)\|^2 \ge 2\mu\{R(\theta) - R(\theta_\star)\}. \tag{127}$$

To verify (127) without leaving the stated stability region, apply strong convexity only to the pair $(\theta, \theta_\star)$:

$$R(\theta_\star) \ge R(\theta) - g(\theta)^\mathsf{T}(\theta - \theta_\star) + \frac{\mu}{2}\|\theta - \theta_\star\|^2.$$

Hence, by Young's inequality,

$$R(\theta) - R(\theta_\star) \le g(\theta)^\mathsf{T}(\theta - \theta_\star) - \frac{\mu}{2}\|\theta - \theta_\star\|^2 \le \frac{\|g(\theta)\|^2}{2\mu},$$

which is (127).

Let $u_t = \mathbb{E}\{R(\theta_t) - R(\theta_\star)\}$. Combining (126) and (127), $u_t \le (1 - c_0\eta_t)u_{t-1} + C_0\eta_t^2$. A standard comparison for regularly varying stepsizes with $0 < \alpha < 1$ gives $u_t = O(\eta_t)$. For completeness, choose $K > C_0/c_0$ and suppose inductively that $u_{t-1} \le K\eta_{t-1}$. Since $\eta_{t-1}/\eta_t = 1 + \alpha/t + O(t^{-2})$ while $\eta_t \asymp t^{-\alpha}$ and $t^{-1} = o(\eta_t)$, for all sufficiently large t, $(1 - c_0\eta_t)K\eta_{t-1} + C_0\eta_t^2 \le K\eta_t$. Enlarging K handles the finite prefix. This proves the first part of (125). Strong convexity also gives $R(\theta) - R(\theta_\star) \ge \frac{\mu}{2}\|\theta - \theta_\star\|^2$, so the second part follows. □

Remark C.43 (Use for quantile regression)**.** *The anchored population check loss risk is locally twice differentiable under the density assumptions, with Hessian A_τ at the target. Positive definiteness of A_τ therefore supplies local strong convexity. A separate stability theorem, or a projected or truncated algorithm with its own rate analysis, is still needed to guarantee localization from an arbitrary initialization; this global issue is separated from the inferential CLT.*

Attainable stability and moment conditions The following scalar examples show how the algorithmic assumptions can be verified without inferring stability from the desired central limit theorem.

Binary logistic regression. Take $X = 1$, $Y \sim \mathrm{Bernoulli}(1/2)$, deterministic initialization, and a clipping interval containing $\mathrm{A} = 1/4$. Let $V(e) = \cosh(e) - 1$. The mean score is $g(e) = \frac{1}{2}\tanh(e/2)$, so $V'(e)g(e) = V(e)/2$. Write the update as $e_t = e_{t-1} + d_t$. Since $|d_t| \le \eta_t/\ell$ and $\cosh(e + d) \le e^{|d|}\cosh(e)$, Taylor's formula gives

$$\mathbb{E}\{V(e_t) \mid \mathcal{F}_{t-1}\} \le V(e_{t-1}) - \frac{\eta_t}{2u}V(e_{t-1}) + C\eta_t^2\{V(e_{t-1}) + 1\}.$$

For small enough η_0, the last coefficient is absorbed into the negative drift. The comparison used in Proposition C.42 yields $\mathbb{E}V(e_t) = O(\eta_t)$. Since $e^2 \le 2V(e)$, (T1) follows. This is a binary-response instance of Corollary C.11; it requires no deterministic confinement and does not establish a general multidimensional logistic rate.

Compact-support quantiles. Let $X = 1$, $\tau = 1/2$, and let Y have a smooth symmetric density supported on $[-b, b]$, positive at zero. Write $D = \eta_0/(2\ell)$ and initialize in $[-b - D, b + D]$. Each update has magnitude at most D. Outside $[-b, b]$ it points inward, so this enlarged interval is invariant. For $g(e) = F_Y(e) - 1/2$, continuity and positivity near zero give $eg(e) \ge ce^2$ on the invariant interval for some $c > 0$. Consequently,

$$\mathbb{E}(e_t^2 \mid \mathcal{F}_{t-1}) \le (1 - 2c\eta_t/u)e_{t-1}^2 + \eta_t^2/(4\ell^2).$$

For sufficiently small η_0, the comparison gives (T1). A slightly larger open interval supplies the tube, and bounded Lipschitz density supplies the envelopes, so both localization and the rate in (Q5) are attained. For bounded-response Huber regression, take the same scalar design, $Y \in [-b, b]$, $c > 2b$, and $\eta_0/\ell \le 1$. From initialization in $[-b, b]$, every update is a convex combination of the current iterate and Y, and the Huber score equals the least squares score throughout this invariant interval.

Functional inference. Take scalar least squares with $X = 1$, Gaussian centered errors, $V_t \in \{-1, 1\}$ and $\mathrm{B}_0 = \mathrm{A} = 1$ inside the clipping interval. Every secant is exactly one, so $\mathrm{B}_t = 1$ and the iterate is a linear Gaussian recursion. With deterministic initialization its mean decays faster than any polynomial and its variance is $O(\eta_t)$. Thus $\mathbb{E}|e_t|^p = O(\eta_t^{p/2})$ for every fixed p, verifying Proposition C.31 and giving a nondegenerate instance of the functional and self-normalized results.

Examples of the exponent q

(i) A pathwise Lipschitz score satisfies (49) with $q = 0$. This includes least squares, logistic, and Huber scores under the stated moments.

(ii) For a shared observation quantile secant, Proposition C.13 shows that $q = 1$ is admissible.

(iii) For independent noisy endpoint scores, Proposition C.14 gives $q = 2$ when the score variance is nondegenerate.

(iv) For another discontinuous estimating equation, q can be found by multiplying the squared jump after division by δ by the probability that the perturbation crosses its discontinuity set.