ONLINE BOOTSTRAP FOR AVERAGED STOCHASTIC APPROXIMATION WITH MOMENTUM

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop an online multiplier bootstrap for inference on parameters estimated by Polyak-Ruppert averaging in two classes of stochastic approximation algorithms: fixed momentum with fixed preconditioning, and dynamic preconditioning without momentum. The procedure updates weighted replicas alongside the original recursion, preserving one-pass computation and avoiding explicit estimation of the asymptotic covariance. Under mean and score regularity and rate conditions for both the original and weighted paths, we show that fixed momentum and fixed preconditioning do not affect the first-order influence function, and the resulting bootstrap remains valid for nonsmooth problems such as quantile regression. For dynamic preconditioning, we establish conditions under which the evolving preconditioner is asymptotically negligible, allowing bootstrap replicas to either reuse the original preconditioner sequence or maintain their own curvature states. This flexibility makes it possible to trade computational cost against how closely the replicas reproduce the original algorithm, without changing first-order inference. The framework supports valid inference with a fixed number of replicas as well as empirical bootstrap quantiles when more replicas are available. We also show that multiple evaluations generated from the same observation should share a common multiplier to preserve the correct covariance structure. Simulations and an application to the UCI Adult data illustrate the finite-sample performance and practical implementation of the proposed approach.

1 INTRODUCTION

Stochastic approximation processes observations recursively (Robbins & Monro, 1951), and Polyak-Ruppert averaging provides a basis for inference (Ruppert, 1988; Polyak & Juditsky, 1992; Kushner & Yin, 2003). Let $g(\theta) = \mathbb{E}\{G(\theta, Z)\}$ be the mean estimating score, where $G : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ and Z is an observation taking values in $\mathcal{Z}$. If $\theta_\star \in \mathbb{R}^d$ solves $g(\theta) = 0$ and $\mathrm{A} = \nabla g(\theta_\star)$ is nonsingular, then, under suitable regularity, $\bar{\theta}_n = n^{-1} \sum_{t=1}^{n} \theta_t$ satisfies

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

Here Q is the score-innovation covariance, $N(0, \mathrm{V})$ the centered normal law with covariance V, and $\Longrightarrow$ convergence in distribution (Toulis & Airoldi, 2017; Duchi & Ruan, 2021). Online inference must recover this law without revisiting the data, and the recursion does not supply its sandwich covariance.

Online inference can estimate the limiting covariance or approximate the sampling law through recursively weighted paths (Chen et al., 2020; Fang et al., 2018). The question here is whether recursive resampling remains valid when the algorithm also carries momentum or curvature states.

Momentum retains past scores, and dynamic preconditioning changes the matrix rescaling future scores, so bootstrap validity must control these perturbed states as well as the original path; a dynamic replica may keep its own matrix state or reuse the original sequence. We study these choices for fixed normalized momentum with fixed nonsingular preconditioning and, separately, predictable dynamic preconditioning without momentum.

Our analysis establishes paired asymptotic linear representations. For centered target innovation ψ_t and an independent mean-one, variance-one primary observation multiplier W_t, the conditions

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.

below give

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) = -\mathrm{A}^{-1} n^{-1/2} \sum_{t=1}^{n} \psi_t + o_p(1),$$

$$\sqrt{n}(\bar{\theta}_n^\star - \theta_\star) = -\mathrm{A}^{-1} n^{-1/2} \sum_{t=1}^{n} W_t \psi_t + o_{\mathbb{P}\otimes\mathbb{P}^\star}(1).$$

Here $\bar{\theta}_n^\star$ averages a weighted replica, $o_p(1)$ is a term converging to zero in probability, and $\mathbb{P} \otimes \mathbb{P}^\star$ is the joint stream-multiplier law. Subtracting gives a conditional multiplier sum with the original limiting covariance; establishing the weighted representation with the additional algorithmic states is the principal task.

The main contributions are as follows.

(i) For fixed normalized momentum with fixed nonsingular preconditioning, each replica weights the incoming score before the momentum filter, preserving one-pass computation; under local score regularity and stability of both paths, the averages share the influence function of averaged SGD, covering smooth and locally nonsmooth scores such as quantile regression without explicit covariance estimation.

(ii) For predictable dynamic preconditioning without momentum, path stability and effective inverse-drift variation remove the preconditioner from the leading influence, so joint replicas with their own curvature states and frozen replicas sharing the original sequence are both valid; coupled through common primary multipliers they are first-order equivalent, and freezing saves curvature computation and storage. Curvature consistency is not needed once stability and variation hold.

(iii) We validate fixed-replica Student intervals (Lam & Wang, 2026), requiring fixed $B \geq 1$, and multivariate F regions (Lam, 2022), requiring $B \geq q$ for nondegenerate q-dimensional contrasts, for both classes. Increasing B consistently approximates conditional quantiles, and conditional uniform integrability adds consistent covariance estimation and covariance-based simultaneous intervals.

1.1 Related work

A limiting Gaussian law must be turned into a computable confidence procedure. Covariance-based inference includes plug-in and batch-means estimators (Chen et al., 2020), fully online iterate-based estimation (Zhu et al., 2023), equal-batch bias correction (Singh et al., 2025), and Hessian-free bias reduction (Wei et al., 2026). Random scaling (Lee et al., 2022) and HiGrad's hierarchical branching and decorrelation (Su & Zhu, 2023) give alternative confidence procedures. Beyond smooth gradients, asymptotic normality and local optimality for nonsmooth stochastic approximation (Davis et al., 2024) are complemented by online covariance estimation (Jiang et al., 2025) and density-free random scaling for quantile regression (Lee et al., 2025); with function evaluations, Jin et al. (2021) study averaged zeroth-order limits and covariance estimation, and Chen et al. (2024) develop Kiefer-Wolfowitz inference, including some nonsmooth losses, with covariance depending on query directions and counts.

Fang et al. (2018) establish conditional bootstrap consistency for Polyak-Ruppert averaged SGD, including quantile regression without explicit density estimation. Quantitative approximations are available for averaged linear stochastic approximation (Samsonov et al., 2024) and smooth nonlinear SGD (Sheshukova et al., 2026). Consistent conditional-quantile approximation uses increasing replica counts, whereas Lam & Wang (2026), building on the cheap-bootstrap principle (Lam, 2022), obtain asymptotically exact scalar intervals from any fixed number $B \geq 1$ of exponentially weighted averaged-SGD paths through t_B calibration and asymptotic independence of the original error and centered replica differences; small B can give wide intervals. With zero momentum, identity preconditioning, and exponential multipliers, our recursion and scalar calibration coincide with their cheap online bootstrap. They also treat offline SGD bootstrap and refined finite-sample bounds under primitive smoothness and convexity assumptions; our extension to the two algorithm classes above is first-order asymptotic, and the multivariate fixed-replica calibration follows Lam (2022).

For momentum and adaptive curvature, the resampling question is whether the extra states remain negligible in the weighted paths, not only in the original estimator. Tang et al. (2023) establish Gaussian limits and averaged-SGD equivalence for smooth strongly convex momentum in their constant-within-run step-size, batch-size, and delayed-averaging regimes. Boyer & Godichon-Baggioni (2023; 2024) study efficiency of stochastic Newton algorithms and weighted averages. An & Huo (2026a) isolate the dynamic-preconditioner contribution and its stabilization threshold; their SA-Adam analysis couples adaptive preconditioning with vanishing momentum gain (An & Huo, 2026b). For randomized Newton solvers, Wang et al. (2026) give last-iterate accelerated-sketching inference through online covariance estimation. Du et al. (2026) establish Gaussian and functional limits for averaged accelerated sketched-Newton iterates, supporting random scaling. These results concern the original estimators; ours establish validity for the weighted replicas and identify when dynamic replicas may share the original preconditioner sequence.

Section 2 defines the online procedures, Section 3 establishes their guarantees, Section 4 evaluates estimation error, coverage, and observation-level weighting, and Section 5 concludes; the appendices give implementation details, model specializations, proofs, additional numerical results with the Adult data illustration, and extensions.

2 METHODOLOGY

2.1 ESTIMATING EQUATIONS AND MULTIPLIER REPLICAS

Throughout, $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced operator norm for matrices, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Let $\theta_\star \in \mathbb{R}^d$ satisfy $g(\theta_\star) = 0$. At iteration t, all paths share a block $\mathcal{U}_t$ containing Z_t and any auxiliary observations or query randomness, and evaluate scores at their own pre-update parameters. Let $\mathcal{F}_{t-1}$ and $\mathcal{F}_{t-1}^{(b)}$ denote the original and replica pre-update histories, respectively.

For either history $\mathcal{H}_{t-1}$, assume the unweighted, possibly approximate score $\widehat{G}_t$ satisfies $\mathbb{E}\{\widehat{G}_t(\theta) \mid \mathcal{H}_{t-1}\} = g(\theta) + b_t(\theta)$ at $\theta_\star$ and at the predictable parameter values used, where b_t is a predictable approximation bias, and define the centered score error $\xi_t(\theta) = \widehat{G}_t(\theta) - g(\theta) - b_t(\theta)$ and its value at the target $\psi_t = \xi_t(\theta_\star)$. For direct scores, $\widehat{G}_t(\theta) = G(\theta, Z_t)$, where $G : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ is measurable and $\mathbb{E}\{G(\theta, Z_t)\} = g(\theta)$; then $b_t \equiv 0$ and $\psi_t = G(\theta_\star, Z_t)$.

Replica multipliers are independent and identically distributed (i.i.d.), independent of the original inputs and across replicas, with $\mathbb{E}W_t = 1$ and $\mathrm{Var}(W_t) = 1$, for example equiprobable $0, 2$. All paths start at the same deterministic θ_0, use $\eta_t = \eta_0 t^{-\alpha}$, and report

$$\bar{\theta}_n = n^{-1} \sum_{t=1}^{n} \theta_t$$

and

$$\bar{\theta}_n^{(b)} = n^{-1} \sum_{t=1}^{n} \theta_t^{(b)}$$

. A superscript $\star$ denotes a generic replica, and $\mathbb{P}^\star$, $\mathbb{E}^\star$, $\mathcal{L}^\star$ denote probability, expectation, and law conditional on the original history $\mathcal{F}_n$. Further conventions and implementation details are in the theory appendix.

2.2 OBSERVATION-LEVEL PERTURBATIONS

Assign one multiplier to each independent observation and reuse it for every evaluation or direction computed from that observation. Distinct independent auxiliary observations receive independent weights, even within one block. For a scalar loss L, query radius $h_t > 0$, and direction U_t, define the central-difference score

$$D_{h_t}(\theta, Z_t, U_t) = \frac{L(\theta + h_t U_t, Z_t) - L(\theta - h_t U_t, Z_t)}{2h_t} U_t,$$

weighted as $W_t D_{h_t}(\theta, Z_t, U_t)$ with the same weight on both loss evaluations. Independent side weights break the cancellation between nearby loss evaluations: loss-level noise divided by h_t gives

variance of order h_t^{-2} under the nondegenerate loss-moment conditions of Proposition C.12, and independent weights for directions sharing one observation delete their cross-covariance terms, with the exact difference in Proposition C.14. These distortions are harmless only when the affected terms vanish or are first-order negligible, and approximation bias must still satisfy the validity conditions.

2.3 Fixed normalized momentum

For fixed $\beta \in [0, 1)$ and nonsingular P, initialize $m_0 = m_0^{(b)} = 0$ and update

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

Weighting precedes the filter, so each perturbation propagates through later states; besides nonsingular P, both paths must be stable.

2.4 Dynamic preconditioning

Without momentum, the original, frozen, and joint parameter updates are

$$\theta_t = \theta_{t-1} - \eta_t \mathrm{P}_t \widehat{G}_t(\theta_{t-1}),$$
$$\theta_t^{F(b)} = \theta_{t-1}^{F(b)} - \eta_t \mathrm{P}_t W_t^{(b)} \widehat{G}_t(\theta_{t-1}^{F(b)}),$$
$$\theta_t^{J(b)} = \theta_{t-1}^{J(b)} - \eta_t \mathrm{P}_t^{(b)} W_t^{(b)} \widehat{G}_t(\theta_{t-1}^{J(b)}).$$

The preconditioners are predictable. For a projected curvature construction, let $\Pi_{\ell,u}$ clip symmetric eigenvalues to $[\ell, u]$, $0 < \ell < u$, and write $\mathcal{S}_{\ell,u} = \{\mathrm{B} = \mathrm{B}^\mathsf{T} : \ell \mathrm{I}_d \preceq \mathrm{B} \preceq u \mathrm{I}_d\}$, where $\preceq$ denotes the positive-semidefinite order. Let $\widehat{\mathrm{H}}_t(\theta)$ be a symmetric curvature observation and $\lambda_t > 0$ the curvature-update gain. Initialize a common deterministic $\mathrm{B}_0 \in \mathcal{S}_{\ell,u}$ and set $\mathrm{P}_1 = \mathrm{P}_1^{(b)} = \mathrm{B}_0^{-1}$. Update

$$\mathrm{B}_t = \Pi_{\ell,u}[\mathrm{B}_{t-1} + \lambda_t\{\widehat{\mathrm{H}}_t(\theta_{t-1}) - \mathrm{B}_{t-1}\}], \qquad \mathrm{P}_{t+1} = \mathrm{B}_t^{-1},$$
$$\mathrm{B}_t^{(b)} = \Pi_{\ell,u}[\mathrm{B}_{t-1}^{(b)} + \lambda_t\{\widehat{\mathrm{H}}_t^{(b)}(\theta_{t-1}^{J(b)}) - \mathrm{B}_{t-1}^{(b)}\}], \quad \mathrm{P}_{t+1}^{(b)} = (\mathrm{B}_t^{(b)})^{-1}.$$

Curvature evaluations use pre-update parameters, so current observations affect only the next preconditioner. Frozen replicas reuse the original time-varying sequence P_t: the preconditioner is shared across replicas rather than held constant over time. A joint curvature observation reuses $W_t^{(b)}$ when the primary record also enters curvature; independent auxiliary observations receive independent weights. The theory appendix gives a score-secant construction, specialized to quantile regression in Corollary C.11.

2.5 Online inference and implementation

For $B \geq 1$ replicas, a fixed contrast $a \in \mathbb{R}^d$, and error probability $\zeta \in (0, 1)$, let $t_{B,p}$ denote the p-quantile of a Student distribution with B degrees of freedom. At a predetermined reporting time, form $\Delta_{n,b} = \bar{\theta}_n^{(b)} - \bar{\theta}_n$ and $T_n^{(b)} = \sqrt{n}\Delta_{n,b}$. For contrast a with positive limiting variance, fixed-B inference uses

$$a^\mathsf{T}\bar{\theta}_n \;\pm\; t_{B,1-\zeta/2}\left\{B^{-1}\sum_{b=1}^{B}(a^\mathsf{T}\Delta_{n,b})^2\right\}^{1/2},$$

a second moment centered at the original estimate with divisor and Student degrees of freedom both B. For $B_n \to \infty$, let $\widehat{q}_p$ be the empirical left p-quantile of $a^\mathsf{T}T_n^{(b)}$; the basic interval is

$$[a^\mathsf{T}\bar{\theta}_n - \widehat{q}_{1-\zeta/2}/\sqrt{n},\; a^\mathsf{T}\bar{\theta}_n - \widehat{q}_{\zeta/2}/\sqrt{n}]$$

. All replicas are initialized with the original recursion and updated from the first observation onward, and the coverage guarantees apply at a prespecified reporting time; Algorithm 1 gives pseudocode. Corollaries C.16 and C.19 and Proposition C.18 justify vector F regions, covariance estimation, and simultaneous inference, with rank and variance fallback rules in Section C.1.5. With fixed B, storage is independent of stream length: dense frozen curvature uses $O(d^2 + Bd)$ memory and one projection/inversion per update, versus $O\{(B+1)d^2\}$ memory and $B+1$ projections/inversions for joint curvature.

3 THEORETICAL RESULTS

Write $\mathrm{A} = \nabla g(\theta_\star)$, Q for the target-innovation covariance, and $\mathrm{V} = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$. The results separate conditional transfer from algorithmic linearization; complete assumptions and verification routes are in Section C.2 and all proofs in Section D.

Conditional transfer. For observed influences ψ_{nt}, Assumption C.1 requires $n^{-1}\sum_t \psi_{nt}\psi_{nt}^{\mathsf{T}} \xrightarrow{p} \mathrm{Q}$, $n^{-1/2}\sum_t \psi_{nt} \Longrightarrow N(0,\mathrm{Q})$, together with multiplier Lindeberg

$$\frac{1}{n}\sum_{t=1}^{n} \|\psi_{nt}\|^2 \mathbb{E}\big[(W_t-1)^2 \mathbf{1}\{|W_t-1|\|\psi_{nt}\| > \varepsilon\sqrt{n}\}\big] \xrightarrow{p} 0 \quad (\varepsilon > 0).$$

These conditions allow triangular-array influences and hold for i.i.d. influences with a finite second moment; Assumption C.2 is the pair of representations displayed in the introduction. Let d_{BL} denote the bounded-Lipschitz distance.

Theorem 3.1. *Under Assumptions C.1 and C.2,*

$$d_{\mathrm{BL}}\left[\mathcal{L}^\star\{\sqrt{n}(\bar\theta_n^\star - \bar\theta_n)\}, N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}})\right] \xrightarrow{p} 0, \tag{2}$$

$$d_{\mathrm{BL}}\left[\mathcal{L}^\star\{\sqrt{n}(\bar\theta_n^\star - \bar\theta_n)\}, \mathcal{L}\{\sqrt{n}(\bar\theta_n - \theta_\star)\}\right] \xrightarrow{p} 0. \tag{3}$$

Fixed momentum. Assumption C.3 requires $1/2 < \alpha < 1$, a local expansion

$$g(\theta_\star + e) = \mathrm{A}e + O(\|e\|^{1+\nu})$$

with $0 < \nu \le 1$ and $\alpha(1+\nu) > 1$, and quadratic-mean continuity

$$\mathbb{E}\|\xi_t(\theta) - \xi_t(\theta_\star)\|^2 \le C\|\theta - \theta_\star\|^\chi, 0 < \chi \le 2$$

. Both paths must have $\mathbb{E}\|\theta_t - \theta_\star\|^2 = O(\eta_t)$ and finite initial second moments. Local hypotheses require globally extended auxiliary paths with uniform bounds, exact agreement after $k_n = o(\sqrt{n})$ with probability tending to one, and uniformly bounded initial-block moments; entering a neighborhood alone is insufficient. Proposition C.4 gives a global Lyapunov route.

Theorem 3.2. *Under Assumption C.3,*

$$\sqrt{n}(\bar\theta_n - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t + o_p(1), \tag{4}$$

$$\sqrt{n}(\bar\theta_n^\star - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} W_t\psi_t + o_{\mathbb{P}\otimes\mathbb{P}^\star}(1). \tag{5}$$

Consequently,

$$d_{\mathrm{BL}}\left[\mathcal{L}^\star\{\sqrt{n}(\bar\theta_n^\star - \bar\theta_n)\}, N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}})\right] \xrightarrow{p} 0, \qquad \mathrm{Q} \textit{ is the matrix in Assumption C.1.} \tag{6}$$

Neither the fixed momentum coefficient β nor the fixed preconditioner P *appears in the first-order influence function.*

Summing the normalized momentum update writes the score sum as weighted parameter increments plus a momentum endpoint, boundary terms that the path rates make negligible at scale $\sqrt{n}$, leaving the same target innovation for both paths; the result is pointwise in fixed $\beta < 1$ and fixed P. Quadratic-mean continuity also permits nonsmooth scores: the quantile-regression specialization in the theory appendix uses local density regularity and a mixed $(2+\nu)$ design moment, together with the separate path-rate requirement. A checkable sufficient route to the path rates is a globally Lipschitz mean g with $g(\theta_\star) = 0$ and a positive-definite matrix R such that

$$e^{\mathsf{T}}\mathrm{R}\mathrm{P}g(\theta_\star + e) \ge \mu e^{\mathsf{T}}\mathrm{R}e, \qquad \mathbb{E}\{\|G(\theta_\star + e, Z_t)\|^2 \mid \mathcal{H}_{t-1}\} \le C(1 + \|e\|^2)$$

for some $\mu > 0$ and every predictable error e; with unbiased scores and finite initial second moments, Proposition C.4 then gives the $O(\eta_t)$ rates for both paths. For a globally strongly convex risk with Lipschitz gradient and positive-definite P, $\mathrm{R} = \mathrm{P}^{-1}$ verifies the drift inequality, while the score-moment bound must still hold; this is a concrete stability route without restricting the theorem to that global class.

Dynamic preconditioning. Assumption C.6 retains the same mean expansion, target-influence conditions, conditional score continuity at predictable states, and both path rates. For each path ϑ_t, approximation bias must satisfy

$$\frac{1}{\sqrt{n}}\sum_t b_t(\theta_\star) \xrightarrow{p} 0, \quad \frac{1}{n}\sum_t \|b_t(\theta_\star)\|^2 \xrightarrow{p} 0, \quad \frac{1}{\sqrt{n}}\sum_t \mathbb{E}\|b_t(\vartheta_{t-1}) - b_t(\theta_\star)\| \to 0.$$

With $\mathrm{M}_t = (\mathrm{P}_t\mathrm{A})^{-1}$, predictable preconditioners must obey $n^{-1/2}\sum_{t=1}^{n-1} t^{\alpha/2}\{\mathbb{E}\|\mathrm{M}_{t+1} - \mathrm{M}_t\|_{\mathrm{F}}^2\}^{1/2} \to 0$, and likewise $\mathrm{M}_t^\star$, with finite initial matrix second moments (Assumption C.7). Local auxiliary paths must preserve the target innovation and bias and satisfy these conditions uniformly, including their own matrix variation; spectral clipping alone does not establish parameter rates.

Theorem 3.3. *Under Assumptions C.6 and C.7, the original and weighted averages satisfy*

$$\sqrt{n}(\bar\theta_n - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^n \psi_t + o_p(1), \tag{7}$$

$$\sqrt{n}(\bar\theta_n^\star - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^n W_t\psi_t + o_{\mathbb{P}\otimes\mathbb{P}^\star}(1). \tag{8}$$

Therefore Theorem 3.1 applies, and the conditional bootstrap law converges to $N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}})$.

For projected joint curvature, Proposition C.9 verifies the weighted variation condition if

$$n^{-1/2}\sum_{t<n} t^{\alpha/2}\lambda_t[1 + \{\mathbb{E}\|\widehat{\mathrm{H}}_t^\star\|_{\mathrm{F}}^2\}^{1/2}] \to 0$$

; in particular, $\lambda_t = O(t^{-\gamma})$ and curvature second moment $O(t^\rho)$, $\rho \geq 0$, suffice when $2\gamma - \rho > 1 + \alpha$. This controls matrix variation only, oracle, bias, and parameter-rate conditions remaining separate, and extends the original-path stabilization argument (An & Huo, 2026a) to perturbed paths.

Theorem 3.4. *Suppose Assumption C.6 holds for the original path and for the joint replica* (22) *and* (23)*. Suppose* $\mathbb{E}\|\mathrm{M}_1\|_{\mathrm{F}}^2 < \infty$*, the original path satisfies* (54)*, and the joint curvature path satisfies* (57)*. Then*

$$d_{\mathrm{BL}}\left[\mathcal{L}^\star\{\sqrt{n}(\bar\theta_n^J - \bar\theta_n)\}, N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}})\right] \xrightarrow{p} 0. \tag{9}$$

If the frozen and joint replicas use the same primary multiplier sequence and the conditions of Corollary C.8 also hold, then

$$\sqrt{n}(\bar\theta_n^J - \bar\theta_n^F) = o_{\mathbb{P}^\star}(1) \tag{10}$$

in $\mathbb{P}$*-probability.*

The cancellation uses $\mathrm{M}_t\mathrm{P}_t = \mathrm{A}^{-1}$, and summation by parts leaves a matrix-increment remainder controlled by the variation condition, so frozen and joint replicas share the leading multiplier sum, saving replica-specific matrix states without ordering finite-sample coverage. For approximate scores with bias $O(h_t^p)$ and $h_t \asymp t^{-\kappa}$, the sufficient undersmoothing condition is $p\kappa > 1/2$ alongside the path and bias conditions above; Corollary C.10 and Proposition C.33 give the associated curvature rates and boundary mean-shift corrections.

Fixed and increasing replica counts.

Theorem 3.5. *Suppose Assumption C.1 holds and Assumption C.2 holds for each of* B *conditionally independent replicas, where* $B \geq 1$ *is fixed. Let* $\mathrm{V} = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$*. Then*

$$\sqrt{n}\begin{pmatrix} \bar\theta_n - \theta_\star \\ \bar\theta_n^{(1)} - \bar\theta_n \\ \vdots \\ \bar\theta_n^{(B)} - \bar\theta_n \end{pmatrix} \Longrightarrow \begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_B \end{pmatrix}, \qquad Z_0, \ldots, Z_B \text{ i.i.d. } N(0, \mathrm{V}). \tag{11}$$

This joint limit yields the Student interval of Corollary C.15 for fixed $B \geq 1$ and the F regions of Corollary C.16 for nondegenerate q-dimensional contrasts with $B \geq q$; increasing B_n consistently approximates conditional quantiles under Corollary C.17. Covariance consistency additionally requires conditional uniform integrability of $\|T_n\|^2$, $T_n = \sqrt{n}(\bar\theta_n^\star - \bar\theta_n)$ (Proposition C.18);

finite-family simultaneous intervals require positive marginal contrast variances, though their joint covariance may be singular (Corollary C.19). Corollaries C.5, C.11, C.20 and C.23 give smooth, quantile, and zeroth-order specializations. The zero-momentum, identity-preconditioner case recovers the weighted SGD bootstrap (Fang et al., 2018), and fixed-replica calibration builds on Lam (2022); Lam & Wang (2026). Relative to original-path results for momentum and dynamic preconditioning (Tang et al., 2023; An & Huo, 2026a), our weighted-path results justify bootstrap inference for both algorithm classes; Corollaries C.21 and C.22 give least-squares and logistic specializations and Section B further theory-level comparisons.

4 Numerical experiments

We label our multiplier bootstrap Proposed and compare it with its SGD special case ($\beta = 0$), compare Proposed-Frozen with Proposed-Joint, and test observation-level weighting. Each validation cell uses 500 independent streams at nominal marginal coverage 95%, with paired errors on shared streams; the numerical appendix gives tuning details, further diagnostics, and the Adult data illustration.

4.1 Proposed momentum bootstrap and SGD

We use five-dimensional linear regression with $Y_t = X_t^\mathsf{T}\theta_\star + \varepsilon_t$, $\theta_\star = (1,-1,1,-1,1)^\mathsf{T}/\sqrt{5}$, independent $\varepsilon_t \sim N(0, 0.25)$, and covariance $\Sigma_{jk} = r^{|j-k|}$ for $r \in \{0, 0.9, 0.98\}$. Covariates are Gaussian or bounded, the latter applying the same covariance factor to independent equiprobable $-1, +1$ innovations, so the families share covariance, target, and noise variance and isolate sensitivity to the covariate distribution; the alternating target probes slowly learned directions under correlation. At $n \in \{2000, 10000\}$ we compare $\beta \in \{0, 0.5, 0.9\}$ with $\mathrm{P} = \mathrm{I}_5$, zero initial states, and full averaging, all methods sharing observations and multiplier draws within each cell, and report full-vector MSE $\mathbb{E}\|\bar\theta_n - \theta_\star\|^2$ and marginal intervals for $e_1^\mathsf{T}\theta_\star$ and the dense contrast $a^\mathsf{T}\theta_\star$, where e_1 is the first coordinate vector and $a = \theta_\star$.

The comparison concerns each algorithm together with its tuning, not the effect of momentum under a common step-size schedule: each method is tuned separately with the same budget of 12 step schedules $\eta_t = \eta_0(t + 100)^{-\alpha}$, $\alpha \in \{0.6, 0.8\}$, $\eta_0 \in \{0.1, 0.3, 1, 3, 10, 30\}$, the offset controlling initial steps without discarding observations. On 80 independent pilot streams, indexed by j, we minimize

$$\frac{1}{80}\sum_{j=1}^{80}\left[\|\bar\theta_{n,j} - \theta_\star\|^2 + \frac{1}{20}\sum_{b=1}^{20}\|\bar\theta_{n,j}^{(b)} - \bar\theta_{n,j}\|^2\right],$$

the replica displacement being centered at the original average on the same stream; this simulation-based oracle tuning uses the true target, does not optimize coverage, and is frozen before validation. Intervals use the fixed-B Student construction in Section 2.5 with $B = 20$ and equiprobable $0, 2$ weights, as does the SGD comparator.

In Table 1, independent covariates give nearly identical MSE and coverage. Under correlation the pattern depends on the covariate distribution: Gaussian designs retain poor dense coverage without an MSE gain from momentum, while the bounded design improves at $r = 0.9$; at $r = 0.98$ substantial bias remains and momentum lowers coverage despite similar MSE. In the bounded design with $r = 0.9$, momentum with a separately selected step schedule can improve both estimation accuracy and interval coverage, at the larger sample size with only a modest increase in interval length (Table 2).

These gains concern the bounded model and specified tuning protocol, not uniform finite-horizon dominance; the numerical appendix reports all selected parameters, contrast biases and lengths, and an estimation-only tuning diagnostic.

4.2 Proposed-Frozen and Proposed-Joint

We generate logistic responses with $d = 10$, Gaussian regressors of covariance $\Sigma_{jk} = 0.3^{|j-k|}$, and $\theta_{\star,j} = (-1)^{j+1}0.35/\sqrt{j}$. The dense contrast uses $a = (1,\dots,1)^\mathsf{T}/\sqrt{10}$; all paths start at zero and use $\eta_t = 0.2t^{-0.7}$.

Table 1: Independently tuned Proposed momentum bootstrap versus its SGD special case. Full-vector MSE and its paired difference are multiplied by 10^3; coverage is for the dense contrast with nominal 95% Student intervals and $B = 20$. Each row uses 500 paired validation runs. The last column reports momentum ($\beta = 0.9$) minus SGD.

r	n	10^3 MSE			Dense coverage			MSE difference versus SGD
		SGD $\beta = 0$	Proposed $\beta = 0.5$	Proposed $\beta = 0.9$	SGD $\beta = 0$	Proposed $\beta = 0.5$	Proposed $\beta = 0.9$	Proposed $\beta = 0.9$
Gaussian covariates								
0	2000	0.746	0.750	0.758	0.954	0.958	0.962	0.012
0.9	2000	82.513	82.755	84.639	0.006	0.002	0.000	2.127
0.98	2000	545.145	545.550	550.246	0.000	0.000	0.000	5.101
0	10000	0.132	0.132	0.132	0.956	0.956	0.952	4×10^{-4}
0.9	10000	7.364	7.375	7.505	0.140	0.136	0.106	0.141
0.98	10000	264.450	264.803	267.341	0.000	0.000	0.000	2.891
Bounded covariates								
0	2000	0.717	0.716	0.724	0.926	0.928	0.936	0.007
0.9	2000	82.881	10.594	10.782	0.000	0.956	0.898	-72.099
0.98	2000	213.841	213.543	217.777	0.336	0.032	0.010	3.936
0	10000	0.137	0.137	0.137	0.940	0.940	0.938	3×10^{-4}
0.9	10000	7.500	1.236	1.255	0.110	0.970	0.956	-6.245
0.98	10000	38.386	38.195	38.581	0.546	0.244	0.206	0.195

Table 2: Marginal coverage and mean interval length for $a = \theta_\star$ and e_1 in the bounded design with $r = 0.9$, using the frozen oracle pilot selections in Table 1. Length is in contrast units. Each row uses 500 paired validation streams and nominal 95% Student intervals with $B = 20$.

n	Method (β)	Dense contrast		First coordinate	
		Coverage	Length	Coverage	Length
2000	SGD (0)	0	0.191	0.260	0.113
2000	Proposed (0.5)	0.956	0.396	0.976	0.189
2000	Proposed (0.9)	0.898	0.277	0.956	0.144
10000	SGD (0)	0.110	0.094	0.716	0.052
10000	Proposed (0.5)	0.970	0.112	0.972	0.058
10000	Proposed (0.9)	0.956	0.098	0.958	0.054

The oracle curvature construction in Section A.2 uses true response probabilities and a diagonal state initialized at I_d and clipped to $[0.05, 20]$; this mechanism experiment does not evaluate feasible curvature estimation. With update gain $\lambda_t = 0.5t^{-\gamma}$, $\gamma = 1$, and curvature-input second moment $O(t^\rho)$, $\rho = 0$, $\Delta_{\mathrm{stab}} = 2\gamma - \rho - (1+\alpha) = 0.3$ satisfies the rate inequality in Theorem 3.4, separately from path stability and localization. Frozen replicas reuse the evolving primary preconditioner, not a matrix held constant in time, joint replicas update their own states, and both use the same primary multipliers. In Table 3, the small coverage differences do not establish an ordering. The coupled distance $\sqrt{n}\|\bar{\theta}_n^{J(b)} - \bar{\theta}_n^{F(b)}\|$, averaged first over replicas and then over 500 streams, is 2.835 at $n = 10^4$ ($B = 199$) and 1.774 at $n = 5 \times 10^4$ ($B = 20$); since both sample size and replica budget change, this does not isolate a sample-size effect or establish convergence. The computational saving is separate: in this diagonal implementation, freezing avoids Bd curvature states with their coordinatewise clipping and reciprocals, the diagonal counterpart of the dense costs in Section C.1.

Table 3: Marginal 95% dense-contrast coverage for Proposed-Frozen and Proposed-Joint with diagonal preconditioning in the $d = 10$ logistic design, with $(\gamma, \rho) = (1, 0)$. Each row uses 500 streams; calibration and replica count B are listed.

Calibration	n	Proposed-Frozen	Proposed-Joint
Empirical ($B = 199$)	10^4	0.944	0.946
Student ($B = 20$)	10^4	0.946	0.950
Student ($B = 20$)	5×10^4	0.956	0.958

4.3 Observation-level weighting diagnostics

For the scalar quadratic loss $L(\theta; Y, C) = \frac{1}{2}(\theta - Y)^2 + C$, central differences use $h \in \{2^{-1}, \ldots, 2^{-7}\}$. Conditional on the two loss evaluations, a common multiplier preserves cancellation of the additive constant, whereas independent side weights introduce variance of order h^{-2}; Panel A of Table 4 reports variance slopes and the loss-shift ratio after adding 100, averaged over the seven radii within each run and then across runs. Across directions, evaluations from one observation share a multiplier to preserve cross-direction covariance, and independent observation blocks use separate weights; Panel B reports signed $(1, 2)$ covariance quantities, not error norms: estimated minus exact covariance under common weights, exact common minus exact independent-direction covariance, and exact independent-block covariance. Section A.3 details the conditioning, reference matrices, and 500-run design.

Table 4: Observation-level weighting diagnostics: means over 500 independent runs. Estimated variances and the common-weight covariance use 5000 exponential multiplier draws. Panel A uses scalar finite differences; panel B uses five two-dimensional contributions. Slopes, ratios and covariances are defined in Section A.3; these are diagnostics, not confidence intervals.

Construction	Diagnostic	Mean
Panel A: finite-difference sides		
Proposed (common side weight)	variance slope	-3.317×10^{-4}
Independent side weights	variance slope	-1.995
Independent side weights	loss-shift variance ratio	2013.340
Panel B: directions and observation blocks		
Proposed (common observation weight)	signed covariance error	-8.770×10^{-3}
Independent direction weights	signed deleted cross-covariance	0.224
Independent observation blocks	conditional off-diagonal covariance	0.024

5 Concluding remarks

Paired asymptotic linearizations establish online multiplier inference for fixed momentum with fixed preconditioning and, separately, for stabilizing dynamic preconditioning, supporting fixed-replica Student and F calibration and increasing-replica conditional inference while accommodating locally nonsmooth scores and approximate stochastic oracles. Observation-level weighting preserves the covariance structure of repeated evaluations from the same sampling unit, and the stabilization theory shows when frozen and jointly evolving preconditioners share the same first-order inference, so frozen replicas can drop replica-specific curvature states and substantially reduce computation and storage.

The numerical results illustrate these mechanisms across momentum, dynamic preconditioning, and multiple-evaluation settings: in the correlated bounded designs, momentum combined with replica-aware tuning can improve both estimation accuracy and interval coverage, while other designs show that such finite-sample gains depend on the data distribution and tuning, underlining the need to account for replica dispersion when selecting algorithmic parameters. Jointly evolving momentum and preconditioning, quantitative coverage guarantees, and dependence-aware resampling remain natural directions for further study.

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. *arXiv preprint arXiv:2604.23498*, 2026a.

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

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996.

Claire Boyer and Antoine 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. doi: 10.1007/s10589-022-00442-3. URL `https://link.springer.com/article/10.1007/s10589-022-00442-3`.

Claire Boyer and Antoine Godichon-Baggioni. Correction: On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 87:705–706, 2024. doi: 10.1007/s10589-023-00511-1.

X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *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.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309–2352, 2017.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *Annals of Statistics*, 52(4):1485–1508, 2024. doi: 10.1214/24-AOS2401. URL `https://arxiv.org/abs/2301.06632`.

Xinchen Du, Elizaveta Rebrova, Michał Dereziński, and Sen Na. Inference for Newton methods with accelerated sketch-and-project via random scaling. *arXiv preprint arXiv:2609.12421*, 2026. URL `https://arxiv.org/abs/2609.12421`.

J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, 49 (1):21–48, 2021.

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.

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

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

Y. Jin, T. Xiao, and K. Balasubramanian. Statistical inference for Polyak–Ruppert averaged zeroth-order stochastic gradient algorithm. *arXiv preprint arXiv:2102.05198*, 2021.

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

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.

H. Lam and Z. Wang. Cheap bootstrap for fast uncertainty quantification of stochastic gradient descent. *Journal of Machine Learning Research*, 27(93):1–42, 2026.

Henry Lam. A cheap bootstrap method for fast inference. *arXiv preprint arXiv:2202.00090*, 2022. URL https://arxiv.org/abs/2202.00090.

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

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast inference for quantile regression with tens of millions of observations. *Journal of Econometrics*, 249:105673, 2025. doi: 10.1016/j.jeconom.2024.105673. URL https://arxiv.org/abs/2209.14502.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

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

J. Præstgaard and J. A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Annals of Probability*, 21(4):2053–2086, 1993.

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

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

S. Samsonov, E. Moulines, Q.-M. Shao, Z.-S. Zhang, and A. Naumov. Gaussian approximation and multiplier bootstrap for Polyak–Ruppert averaged linear stochastic approximation with applications to TD learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

X. Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489): 218–235, 2010.

M. Sheshukova, S. Samsonov, D. Belomestny, E. Moulines, Q.-M. Shao, Z.-S. Zhang, and A. Naumov. Gaussian approximation and multiplier bootstrap for stochastic gradient descent. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1378–1386. PMLR, 2026. URL https://proceedings.mlr.press/v300/sheshukova26a.html.

R. Singh, A. Shukla, and D. Vats. On the utility of equal batch sizes for inference in stochastic gradient descent. *Journal of Machine Learning Research*, 26(258):1–41, 2025.

J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

W. J. Su and Y. Zhu. HiGrad: Uncertainty quantification for online learning and stochastic approximation. *Journal of Machine Learning Research*, 24(124):1–53, 2023.

K. Tang, W. Liu, Y. Zhang, and X. Chen. Acceleration of stochastic gradient descent with momentum by averaging: Finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2023.

P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Annals of Statistics*, 45(4):1694–1727, 2017.

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

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

H. Wang, X. Du, and S. Na. Inference of online newton methods with Nesterov's accelerated sketching. *arXiv preprint arXiv:2604.23436*, 2026. URL https://arxiv.org/abs/2604.23436. Accepted at the International Conference on Machine Learning (ICML), 2026.

Z. Wei, W. Zhu, J. Lyu, and W. B. Wu. Refining covariance matrix estimation in stochastic gradient descent through bias reduction. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3565–3573. PMLR, 2026. URL https://proceedings.mlr.press/v300/wei26c.html.

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 NUMERICAL RESULTS

The independently tuned comparison uses six scenarios in each of two covariate families, with 500 validation streams shared across three momentum values in every scenario, totaling 6000 streams. Its estimation-only diagnostic uses 3000 separate streams. The dynamic-preconditioning, multiple-evaluation, and finite-difference designs comprise 12 additional cells with 500 streams each. Coverage is evaluated against 0.95 and includes every attempted stream. Fixed-replica scalar intervals use Student calibration; empirical intervals use left bootstrap quantiles. Table 5 summarizes the retained designs.

Table 5: Simulation designs, with 500 streams per cell.

Design	Cells	Streams	Principal factors	Main diagnostic
Tuned momentum	12	6000	n, r, β, covariate family	MSE, coverage, interval length
Estimation-only tuning	6	3000	n, r, β, step schedule	replica dispersion
Dynamic preconditioning	7	3500	γ, ρ, frozen/joint	variation and coupled distance
Multiple evaluations	2	1000	side/direction weights	variance slope, covariance
Finite-difference bias	3	1500	$\kappa \in \{0.200, 0.250, 0.300\}$	paired root-n shift

A.1 INDEPENDENTLY TUNED MOMENTUM: IMPLEMENTATION AND UNCERTAINTY

Each covariate family in Section 4.1 has six scenarios with independent Gaussian errors. Write $\mathrm{L}\mathrm{L}^\intercal = \Sigma$ for the lower-triangular Cholesky factor. We generate $X_t = \mathrm{L}U_t$, where the coordinates of U_t are independent standard normals or equiprobable $-1, +1$ variables. The two scalar targets are $e_1^\intercal \theta_\star$ and $a^\intercal \theta_\star$, where $a = \theta_\star$ has unit norm. Mean squared error (MSE) refers to $\mathbb{E}\|\bar{\theta}_n - \theta_\star\|^2$ over the full vector; coverage and length are marginal for each contrast, not simultaneous. Every average includes all post-update iterates. There is no tail averaging, burn-in deletion, or recentering by a batch fit. The weighted score enters before the momentum filter, and each replica evaluates the score at its own parameter. Numerical parity with an independent R implementation was checked for $\beta \in \{0, 0.5, 0.9\}$.

For each scenario and momentum value, we minimize the pilot criterion in Section 4.1. All methods use the same grid and pilot streams. The chosen parameters are then held fixed; the 500 validation streams have disjoint seeds from all development runs. The offset schedule is asymptotically proportional to $t^{-\alpha}$, but the experiment evaluates its finite-horizon behavior and does not verify the path-rate assumptions. The criterion controls an empirical tradeoff between bias and path dispersion; a finite pilot does not certify stability in the tails.

Table 6: Gaussian design: frozen pilot selections and paired full-vector MSE differences from $\beta = 0$. Differences are multiplied by 10^3; negative values favor momentum. Each row uses 500 paired validation streams. SGD uses $\beta = 0$; Proposed uses $\beta \in \{0.5, 0.9\}$.

r	n	β	α	η_0	10^3 paired MSE difference
0	2000	0	0.8	3	—
0	2000	0.5	0.8	3	0.004
0	2000	0.9	0.8	3	0.012
0.9	2000	0	0.6	1	—
0.9	2000	0.5	0.6	1	0.242
0.9	2000	0.9	0.6	1	2.127
0.98	2000	0	0.6	1	—
0.98	2000	0.5	0.6	1	0.404
0.98	2000	0.9	0.6	1	5.101
0	10000	0	0.8	3	—
0	10000	0.5	0.8	3	0
0	10000	0.9	0.8	3	4×10^{-4}
0.9	10000	0	0.6	1	—
0.9	10000	0.5	0.6	1	0.012
0.9	10000	0.9	0.6	1	0.141
0.98	10000	0	0.6	1	—
0.98	10000	0.5	0.6	1	0.353
0.98	10000	0.9	0.6	1	2.891

Table 7: Gaussian design: marginal Student coverage (C), mean interval length (L), and bias for the first coordinate and dense contrast. Each row uses 500 streams; $B = 20$. Nominal coverage is 95%; length and bias are in contrast units. SGD uses $\beta = 0$; Proposed uses $\beta \in \{0.5, 0.9\}$.

r	n	β	First coordinate			Dense contrast		
			C	L	Bias	C	L	Bias
0	2000	0	0.950	0.053	-0.003	0.954	0.054	-0.007
0	2000	0.5	0.948	0.053	-0.002	0.958	0.054	-0.007
0	2000	0.9	0.952	0.053	-0.002	0.962	0.055	-0.007
0.9	2000	0	0.390	0.127	-0.073	0.006	0.212	-0.266
0.9	2000	0.5	0.370	0.124	-0.073	0.002	0.206	-0.266
0.9	2000	0.9	0.344	0.124	-0.075	0.000	0.202	-0.269
0.98	2000	0	0.006	0.186	-0.237	0.000	0.197	-0.715
0.98	2000	0.5	0.004	0.179	-0.237	0.000	0.188	-0.716
0.98	2000	0.9	0.000	0.170	-0.238	0.000	0.179	-0.719
0	10000	0	0.968	0.022	0.000	0.956	0.022	-0.001
0	10000	0.5	0.966	0.022	0.000	0.956	0.022	-0.001
0	10000	0.9	0.968	0.022	0.000	0.952	0.022	-0.001
0.9	10000	0	0.714	0.054	-0.019	0.140	0.100	-0.075
0.9	10000	0.5	0.716	0.053	-0.019	0.136	0.099	-0.075
0.9	10000	0.9	0.706	0.053	-0.019	0.106	0.098	-0.076
0.98	10000	0	0.002	0.116	-0.144	0.000	0.168	-0.491
0.98	10000	0.5	0.000	0.112	-0.144	0.000	0.162	-0.491
0.98	10000	0.9	0.000	0.110	-0.145	0.000	0.157	-0.494

Table 8: Bounded design: frozen pilot selections and paired full-vector MSE differences from $\beta = 0$. Differences are multiplied by 10^3; negative values favor momentum. Each row uses 500 paired validation streams. SGD uses $\beta = 0$; Proposed uses $\beta \in \{0.5, 0.9\}$.

r	n	β	α	η_0	10^3 paired MSE difference
0	2000	0	0.8	3	—
0	2000	0.5	0.8	3	-8×10^{-4}
0	2000	0.9	0.8	3	0.007
0.9	2000	0	0.6	1	—
0.9	2000	0.5	0.6	3	-72.287
0.9	2000	0.9	0.6	3	-72.099
0.98	2000	0	0.6	3	—
0.98	2000	0.5	0.6	3	-0.298
0.98	2000	0.9	0.6	3	3.936
0	10000	0	0.8	3	—
0	10000	0.5	0.8	3	1×10^{-4}
0	10000	0.9	0.8	3	3×10^{-4}
0.9	10000	0	0.6	1	—
0.9	10000	0.5	0.6	3	-6.264
0.9	10000	0.9	0.6	3	-6.245
0.98	10000	0	0.6	3	—
0.98	10000	0.5	0.6	3	-0.191
0.98	10000	0.9	0.6	3	0.195

Table 9: Bounded design: marginal Student coverage (C), mean interval length (L), and bias for the first coordinate and dense contrast. Each row uses 500 streams; $B = 20$. Nominal coverage is 95%; length and bias are in contrast units. SGD uses $\beta = 0$; Proposed uses $\beta \in \{0.5, 0.9\}$.

r	n	β	First coordinate			Dense contrast		
			C	L	Bias	C	L	Bias
0	2000	0	0.956	0.051	-0.003	0.926	0.051	-0.008
0	2000	0.5	0.958	0.051	-0.003	0.928	0.051	-0.008
0	2000	0.9	0.962	0.051	-0.003	0.936	0.052	-0.008
0.9	2000	0	0.260	0.113	-0.073	0.000	0.191	-0.266
0.9	2000	0.5	0.976	0.189	-0.018	0.956	0.396	-0.064
0.9	2000	0.9	0.956	0.144	-0.018	0.898	0.277	-0.065
0.98	2000	0	0.922	0.489	-0.119	0.336	0.789	-0.420
0.98	2000	0.5	0.720	0.309	-0.119	0.032	0.482	-0.420
0.98	2000	0.9	0.636	0.287	-0.120	0.010	0.439	-0.426
0	10000	0	0.952	0.021	-0.001	0.940	0.021	-0.002
0	10000	0.5	0.952	0.021	-0.001	0.940	0.021	-0.002
0	10000	0.9	0.950	0.021	-0.001	0.938	0.021	-0.002
0.9	10000	0	0.716	0.052	-0.019	0.110	0.094	-0.076
0.9	10000	0.5	0.972	0.058	-0.003	0.970	0.112	-0.013
0.9	10000	0.9	0.958	0.054	-0.003	0.956	0.098	-0.013
0.98	10000	0	0.902	0.188	-0.043	0.546	0.385	-0.169
0.98	10000	0.5	0.832	0.138	-0.044	0.244	0.260	-0.169
0.98	10000	0.9	0.798	0.133	-0.043	0.206	0.250	-0.171

Every attempt is retained, and a numerical failure would count as noncoverage. No nonfinite-state failure occurred in either validation family or in the estimation-only diagnostic; large but finite errors were not discarded.

Why the weighted paths enter tuning. An estimation-only tuning diagnostic uses the same six models and 12 schedules, 40 independent pilot streams, and 500 separate validation streams. It minimizes only original-path MSE. This can favor steps whose weighted paths have very large dispersion even when the original average is reasonably accurate. For example, at $r = 0.9, n = 2000$, dense-contrast mean interval lengths are 152.800, 681.871, and 4.297 for $\beta = 0, 0.5, 0.9$, with coverage 1.000, 1.000, and 0.998. Such coverage is not evidence of efficient uncertainty quantification. At $r = 0.98, n = 2000$, the zero-momentum MSE is 7.631, revealing sensitivity to rare large errors. The replica-aware criterion was specified after this diagnostic, using new pilot and validation seeds. The bounded-design family assesses covariate-tail sensitivity after the Gaussian pilot selected conservative steps for every momentum value; it has separate pilot and validation seeds. No scenario was removed and no schedule was changed after its validation. The complete diagnostic summaries are in Table 10.

Table 10: Estimation-only tuning diagnostic in Gaussian designs. Each row uses 500 independent validation streams. C_a and L_a denote dense-contrast coverage and mean length. Intervals use nominal 95% Student calibration with $B = 20$; length is in contrast units and full-vector MSE is multiplied by 10^3. SGD uses $\beta = 0$; Proposed uses $\beta \in \{0.5, 0.9\}$.

r	n	β	α	η_0	10^3 MSE	C_a	L_a
0	2000	0	0.8	3	0.749	0.942	0.054
0	2000	0.5	0.6	3	0.783	0.994	0.095
0	2000	0.9	0.8	3	0.760	0.942	0.054
0.9	2000	0	0.6	3	14.680	1.000	152.800
0.9	2000	0.5	0.8	10	19.122	1.000	681.871
0.9	2000	0.9	0.8	10	10.731	0.998	4.297
0.98	2000	0	0.6	3	7631.111	0.998	14437.396
0.98	2000	0.5	0.6	3	230.477	0.980	96.637
0.98	2000	0.9	0.6	3	231.738	0.820	4.209
0	10000	0	0.8	3	0.129	0.942	0.022
0	10000	0.5	0.8	3	0.130	0.940	0.022
0	10000	0.9	0.8	3	0.130	0.936	0.022
0.9	10000	0	0.6	3	1.386	1.000	263.809
0.9	10000	0.5	0.6	3	1.313	1.000	1.531
0.9	10000	0.9	0.8	10	1.380	1.000	0.703
0.98	10000	0	0.6	3	50.831	0.998	2320.254
0.98	10000	0.5	0.6	3	40.426	0.992	66.457
0.98	10000	0.9	0.6	3	40.509	0.914	1.484

A.2 Proposed-Frozen and Proposed-Joint: implementation

The logistic model has $X_t \sim N(0, \Sigma)$, $\Sigma_{jk} = 0.3^{|j-k|}$, and $Y_t \mid X_t \sim \text{Bernoulli}(p_t)$ with $p_t = \{1 + \exp(-X_t^\top \theta_\star)\}^{-1}$ and $\theta_{\star,j} = (-1)^{j+1} 0.35/\sqrt{j}$. The primary design uses $d = 10$, $a = (1, \ldots, 1)^\top/\sqrt{d}$, zero parameter initialization, and $\mathrm{B}_0 = \mathrm{I}_d$. The independently selected step scale is $\eta_0 = 0.2$, with $\eta_t = \eta_0 t^{-0.7}$.

The curvature state is diagonal. Its input has entries $X_{tj}^2 p_t(1 - p_t) + 0.15 + 0.15 t^{\rho/2} U_{tj}$, where U_t is independent and uniform on the unit sphere. This oracle mechanism experiment uses the true p_t and does not evaluate a feasible Hessian estimator at the current iterate. With $\lambda_t = 0.5 t^{-1}$, original and joint updates clip each diagonal entry to $[0.05, 20]$, using respectively the input and its observation-weighted version. Frozen replicas use the evolving original diagonal inverse; joint replicas use their own. Within each joint replica, parameter and curvature updates share the same equiprobable $0, 2$ multiplier. The input second moment is $O(t^\rho)$ and $\gamma = 1$ controls the gain decay. Taking $\rho \in \{0, 0.3, 0.6\}$ gives

$$\Delta_{\text{stab}} = 2\gamma - \rho - (1 + \alpha) \in \{0.3, 0, -0.3\}.$$

The positive and boundary settings use $d = 10$; the negative setting uses $d = 30$, with the same coordinatewise target formula and normalized all-ones contrast. The strict positive cell satisfies the

sufficient variation-rate inequality; equality and the negative cell are stress tests. These cell labels do not verify the separate path-stability or localization hypotheses. A biased but slowly varying curvature target checks that curvature consistency is not required once variation and parameter stability hold.

Table 11: Marginal 95% dense-contrast coverage for Proposed-Frozen and Proposed-Joint with diagonal dynamic preconditioning, using 500 streams per row. The regime identifies the rate margin; calibration and replica count are listed.

Regime	Calibration	n	Proposed-Frozen	Proposed-Joint
$\Delta_{\mathrm{stab}} = 0.300$	Empirical (199)	10000	0.944	0.946
$\Delta_{\mathrm{stab}} = 0.300$	Student (20)	10000	0.946	0.950
$\Delta_{\mathrm{stab}} = 0.300$	Student (20)	50000	0.956	0.958
Stress: $\Delta_{\mathrm{stab}} = 0$	Student (20)	50000	0.930	0.936
Stress: $\Delta_{\mathrm{stab}} = -0.300$	Student (5)	50000	0.952	0.958

In the positive-margin setting, the coupled diagnostic is the mean of $\sqrt{n}\|\bar{\theta}_n^{J(b)} - \bar{\theta}_n^{F(b)}\|$ over all maintained replicas within a stream, followed by the mean over 500 streams. It decreases from 2.835 at $n = 10000$ ($B = 199$) to 1.774 at $n = 50000$ ($B = 20$). Both the sample size and replica budget change; the comparison does not isolate a sample-size effect or establish convergence to zero. The boundary and below-boundary cells happen to remain well calibrated in this particular model. Their favorable behavior illustrates why the sufficient rate boundary must not be interpreted as a pointwise necessity.

A.3 Proposed observation-level weighting

For panel A of Table 4, each of 500 independent runs draws $\theta \sim N(0.5, 0.01)$, $Y \sim N(1.1, 0.01)$, and $C \sim N(3, 1)$ independently, then fixes these values. Set $L_\pm = \frac{1}{2}(\theta \pm h - Y)^2 + C$ for $h \in \{2^{-1}, \ldots, 2^{-7}\}$. Conditional on (θ, Y, C), the centered perturbations are $(W-1)(L_+ - L_-)/(2h)$ and $\{(W_+ - 1)L_+ - (W_- - 1)L_-\}/(2h)$. The weights are independent $\mathrm{Exp}(1)$ draws, with W shared across sides in the first construction. Their exact conditional variances are $(L_+ - L_-)^2/(4h^2)$ and $(L_+^2 + L_-^2)/(4h^2)$.

Each conditional variance is estimated from 5000 multiplier draws. Within a run, log estimated variance is regressed on $\log h$ over $h = 2^{-2}, \ldots, 2^{-7}$ for both slopes. The loss-shift diagnostic averages the seven ratios of the estimated independent-side variances after and before adding 100 to both losses, with paired multiplier draws. The table averages each diagnostic over 500 runs.

For panel B, each run fixes five vectors $D_j \in \mathbb{R}^2$ with independent standard normal entries. The centered common-weight perturbation is $(W-1)\sum_j D_j$, while the independent-direction perturbation is $\sum_j (W_j - 1)D_j$. Conditional on these vectors, their reference covariance matrices are $\mathrm{C}_{\mathrm{com}} = (\sum_j D_j)(\sum_j D_j)^{\mathsf{T}}$ and $\mathrm{C}_{\mathrm{ind}} = \sum_j D_j D_j^{\mathsf{T}}$. The common-weight signed error is $[\widehat{\mathrm{C}}_{\mathrm{com}} - \mathrm{C}_{\mathrm{com}}]_{12}$, with $\widehat{\mathrm{C}}_{\mathrm{com}}$ estimated from 5000 draws; the deleted cross-covariance is $[\mathrm{C}_{\mathrm{com}} - \mathrm{C}_{\mathrm{ind}}]_{12}$. For independent blocks, a separate set of five standard normal vectors $\widetilde{D}_j$ gives the exact conditional covariance $[\sum_j \widetilde{D}_j \widetilde{D}_j^{\mathsf{T}}]_{12}$. Its expectation over blocks is zero, but its realized value need not be. All three quantities are signed and averaged over 500 runs; the last is not a covariance-estimation error.

The slopes in Table 4 recover the exact orders: essentially zero for Proposed and -2 for independent side weights. Adding 100 to the loss, which leaves the gradient unchanged, multiplies the invalid conditional variance by more than 2000 on average. Directionwise weights likewise remove cross-direction terms when the directions share a record, but are appropriate when the directions are attached to independent observation blocks. These calculations are primarily algebra and implementation checks; they do not rely on asymptotic coverage.

A.4 THE FINITE-DIFFERENCE BIAS BOUNDARY

A cubic smooth benchmark gives central-difference bias of order $p = 2$. We used $h_t = h_0 t^{-\kappa}$ with $\kappa \in \{0.20, 0.25, 0.30\}$, placing $p\kappa$ below, at, and above $1/2$. The analytic correction and coupled Richardson construction use the same observation multiplier at both scales. All versions use Proposed observation-level bootstrap weighting; the correction labels distinguish treatments of score bias.

Table 12: Finite-difference bias diagnostics for Proposed. The paired shift is the mean root-n analytic-minus-uncorrected contrast. Coverage uses nominal 95% empirical $B = 199$ intervals and 500 streams per cell; length is in contrast units.

κ	Paired shift	Uncorrected cov.	Analytic cov.	Mean length (analytic)
0.200	0.129	0.938	0.944	0.019
0.250	0.059	0.954	0.954	0.019
0.300	0.029	0.946	0.942	0.019

The paired shifts decrease across the boundary, consistently with the predicted h^2 bias term. The three plotted correction curves overlap at the displayed precision. For a one-dimensional cubic loss with unit query direction, central differencing adds its third derivative times $h^2/6$ to the derivative; subtracting this term agrees with Richardson at the score level. Ordinary undersmoothing retains the term at a smaller positive bandwidth, and a correction applied after averaging need not equal a recursion run with corrected scores. The plotted agreement therefore does not establish equality of the algorithms. Coverage is close to 0.95 for both versions, without a consistent gain from analytic correction across the three settings. Centered inference retains the residual-bias and stability conditions following Proposition C.33.

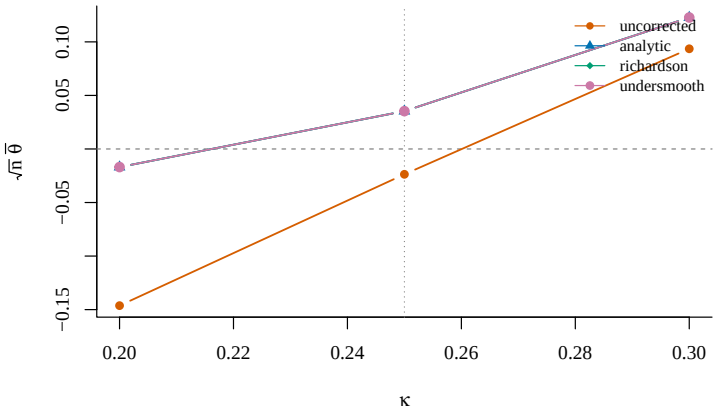


Figure 1: Root-n finite-difference bias diagnostics for Proposed across $\kappa = 0.20, 0.25, 0.30$. The paired contrast in Table 12 is consistent with a leading h^2 component. The correction curves overlap at the displayed precision; this does not establish identity of the implemented algorithms.

A.5 UCI ADULT DETAILS

The real-data analysis illustrates the fixed-momentum, fixed-preconditioner procedure, not the dynamic-preconditioning theorem. We use the UCI Adult data (Becker & Kohavi, 1996), which contain 48842 extracted 1994 Census records and are distributed under CC BY 4.0 with DOI 10.24432/C5XW20. UCI documents the extraction restrictions `AAGE` $>$ 16, `AGI` $>$ 100, `AFNLWGT` $>$ 1, and `HRSWK` $>$ 0. The analysis unit is one UCI record. For this illustration we

treat the final records as an exchangeable i.i.d. working stream; this is a modeling assumption rather than a Census sampling-design assertion.

The binary response is whether recorded annual income exceeds \$50000. The low-dimensional working logistic model uses fixed-unit transforms of age, `education-num`, `hours-per-week`, capital gain and loss, sex, and marital status. No outcome-dependent filtering or imputation is used, and `fnlwgt` is not used as a survey weight. The official train and test files are pooled, cleaned, and shuffled. The first 5000 records initialize the coefficient and a curvature preconditioner for tuning; the next 5000 select a common step schedule for $\beta \in \{0, 0.8\}$ by a worst-over-momentum score-residual rule; and the remaining 38842 records are untouched until the final analysis. The coefficient and fixed preconditioner are then refitted on all 10000 preliminary records. Conditional on this split, the initialization, preconditioner, and schedule $\eta_t = 0.1t^{-0.65}$ are fixed before final-stream inference.

The inferential target is the coefficient $\theta_\star$ solving the unweighted working-distribution logistic estimating equation for an Adult-like record. It is not the deterministic root of the final empirical score. The batch GLM fitted to the 38842 final records is instead a computational benchmark. Both momentum paths use the same record order, fixed preconditioner, $B = 999$ bounded multipliers, and common multiplier draws. The final score-mean norms were 1.490×10^{-2} and 1.530×10^{-2}, and the Polyak-Ruppert estimates were within 1.380 batch standard errors on the two primary contrasts.

Table 13: Adjusted odds ratios in the UCI Adult working model. Multiplier intervals use empirical $B = 999$ calibration; plug-in and batch columns give normal intervals.

Contrast	β	Multiplier	Online plug-in	Batch GLM
Education code $+1$	0	1.457 (1.431, 1.476)	1.457 (1.437, 1.478)	1.444 (1.424, 1.464)
Education code $+1$	0.8	1.458 (1.437, 1.475)	1.458 (1.437, 1.478)	1.444 (1.424, 1.464)
Recorded hours $+10$	0	1.379 (1.336, 1.419)	1.379 (1.342, 1.418)	1.368 (1.331, 1.406)
Recorded hours $+10$	0.8	1.380 (1.341, 1.419)	1.380 (1.343, 1.419)	1.368 (1.331, 1.406)

The education contrast is a one-unit higher value of UCI's ordered numeric code under a linear logistic specification; it is not a homogeneous educational intervention. The hours contrast compares records differing by 10 in `hours-per-week`, holding the included covariates fixed.

The close $\beta = 0$ and $\beta = 0.8$ results show numerical stability of these two contrasts under the chosen, pilot-selected fixed-momentum paths. Under the working model, higher recorded education code and weekly hours are positively associated with the odds of income above \$50000 after adjustment for the included covariates. These historical-record associations are not causal effects, fairness assessments, or design-based estimates for the U.S. population.

B Additional discussion

A first extension is a recursion with constant momentum and a dynamic preconditioner. If $m_t = \beta m_{t-1} + (1-\beta)G(\theta_{t-1}, Z_t)$ and $\theta_t = \theta_{t-1} - \eta_t \mathrm{P}_t m_t$, the fixed-momentum transformation produces the additional term

$$\frac{\beta}{1-\beta}(\eta_{t-1}\mathrm{P}_{t-1} - \eta_t \mathrm{P}_t)m_{t-1}.$$

A valid original and bootstrap linearization therefore requires a joint stability theorem for $(\theta_t, m_t, \mathrm{P}_t)$ and a summability bound for products of matrix increments and momentum states. At the stabilization boundary, the matrix remainder may survive at first order. In models where it does, a frozen bootstrap may be insufficient, and a joint augmented-state construction would need to reproduce the matrix or curvature randomness contributing to that first-order term. The analysis of SA-Adam (An & Huo, 2026b) provides a starting point, but it uses a different momentum schedule from the constant-β recursion considered here.

A second extension is quantitative or non-i.i.d. inference. Finite-sample coverage bounds would have to combine Gaussian approximation, replica approximation, algorithmic linearization, dynamic-matrix variation, and finite-replica calibration; existing results for linear stochastic approximation and smooth SGD (Samsonov et al., 2024; Sheshukova et al., 2026; Lam & Wang, 2026)

provide benchmarks but do not control these terms for the stateful recursions above. Serially correlated influence sequences generally require resampling that reproduces their long-run covariance, such as block or dependent multipliers (Shao, 2010), with each block retaining all evaluations from its sampling units. Martingale-difference dependence can still permit independent multipliers when the original CLT, empirical covariance, multiplier Lindeberg condition, and paired representations are established. Growing-dimensional contrasts would require a uniform linear representation, high-dimensional Gaussian comparison and anti-concentration (Chernozhukov et al., 2017), together with dimension-uniform control of the preconditioner remainder. These extensions change the probability approximation itself and cannot be obtained by modifying only the reporting step of Algorithm 1.

B.1 Assumptions

The assumptions serve different purposes. Some are probability conditions needed for central limit or bootstrap approximations. Others are high-level stability conditions for a particular recursion. We do not claim that the stability conditions are necessary, and they should be replaced by model-specific drift arguments whenever possible. Below, we explain the weak form used in the theorem, how each condition can be checked, and which stronger conditions are not required.

The multiplier normalization in (12) has three roles. Mean one preserves the estimating equation, variance one reproduces the scale of the influence covariance, and independence across replicas permits empirical conditional inference. Two-point, exponential, or other i.i.d. weights can be used after centering and normalization. This is the standard weighted-bootstrap normalization (see Præstgaard & Wellner, 1993; van der Vaart & Wellner, 1996; Fang et al., 2018). The transfer theorem does not require bounded or nonnegative weights, or moments above order two.

The influence conditions in Assumption C.1 provide the minimal ingredients for the conditional multiplier limit: empirical covariance convergence, a Lindeberg condition, and a central limit theorem. For i.i.d. influences, a finite second moment is enough by Lemma C.27. For martingale arrays, these conditions can be checked through conditional variance convergence and a martingale Lindeberg theorem (see Hall & Heyde, 1980). A uniform $2 + \delta$ moment or a largest-term Lyapunov bound is sufficient, but neither is assumed.

The paired linear representations in Assumption C.2 require the original and weighted averages to have the same first-order map. This is needed because the transfer theorem cannot determine the effect of an algorithmic state by itself. The required representations are proved for momentum and dynamic preconditioning in Theorems 3.2 and 3.3. Fang et al. (2018) treat the ordinary SGD case. Once an algorithm-specific theorem applies, the paired representation is an organizing condition rather than an additional model assumption.

The step-size and local mean-expansion conditions in (M1)/(D1) control two remainders. The restriction $\alpha < 1$ makes the boundary terms in the Abel sum negligible, while $\alpha(1 + \nu) > 1$ makes the accumulated nonlinear drift remainder $o(1)$ at the root-n scale. A locally Hölder derivative of g gives the stated mean-expansion remainder; a locally Lipschitz derivative corresponds to $\nu = 1$. These are standard local conditions for averaged stochastic approximation and M-estimation (see Ruppert, 1988; Polyak & Juditsky, 1992; Kushner & Yin, 2003; van der Vaart, 1998). The linearization does not require a globally bounded sample Hessian or global strong convexity.

The second-moment and quadratic-mean continuity conditions in (M2)/(D2) make score variation along the estimated path negligible in mean square. This is stochastic equicontinuity in the norm used by the recursion. An L^2 random Lipschitz envelope gives exponent $\chi = 2$. Indicator scores can have a smaller exponent, obtained through a crossing-probability argument as in Corollary C.5. For raw least squares, Corollary C.21 uses an L^p martingale argument with $p = 1 + \delta/2$ and only $\mathbb{E}\|X\|^{2+\delta} < \infty$, so a fourth design moment is not needed for the first-order representation. These conditions do not require every sample score to be differentiable or Lipschitz continuous.

The parameter-rate and localization conditions in (M3), (M4), and (D4) use $\mathbb{E}\|e_t\|^2 = O(\eta_t)$ to control endpoint terms and turn the local drift and score moduli into summable remainders. A one-step Lyapunov drift gives this rate in Proposition C.31; related arguments are standard (Kushner & Yin, 2003; Moulines & Bach, 2011; Tang et al., 2023). Localization is needed only when the primitive conditions are local. We do not claim that spectral clipping alone implies the rate. Global assumptions can replace localization, but they are stronger.

The oracle-bias condition (D3) is needed because a centered multiplier law reproduces stochastic variation but does not remove a deterministic root-n bias. Local bias variation also enters the weighted score remainder. Exact scores have zero bias. When $b_t(\theta_\star) = O(h_t^p)$, root-$n$ undersmoothing requires $n^{-1/2}\sum_t h_t^p \to 0$; Section C.8 gives the boundary and corrections. Uniform differentiability of b_t is not required, since the theorem uses only the along-path absolute-sum condition in (52).

The predictability, nonsingularity, and effective-inverse initialization requirements in Assumption C.7 support the Abel argument. Predictability preserves the martingale structure, nonsingularity permits the exact identity $\mathrm{M}_t\mathrm{P}_t = \mathrm{A}^{-1}$, and a finite initial second moment anchors the increment-based control. Projecting a symmetric curvature state onto $\mathcal{S}_{\ell,u}$ is one simple way to meet these requirements. More general preconditioners are allowed when the same properties are checked directly (compare An & Huo, 2026a). The first-order theorem does not require convergence of P_t or of a curvature state to A^{-1}.

The weighted matrix variation in (54) to (55) directly bounds the matrix-increment part of the Abel sum. Polynomial increments of order $t^{-\beta_{\mathrm{P}}}$ are sufficient when $\beta_{\mathrm{P}} > (1+\alpha)/2$. The decoupled lower construction (An & Huo, 2026a) shows that these rate hypotheses are sharp, not that they are necessary for every coupled algorithm. The main theorem does not assume a polynomial rate, and non-polynomial paths can be checked directly from the sum. Curvature consistency alone does not imply the required rate.

For a projected joint curvature state, (57) makes replica-specific curvature updates first-order negligible while preserving their finite-sample effect. Since spectral projection is nonexpansive, verification reduces to checking the gain and the second moment of the curvature observation. Under polynomial bounds, this gives $2\gamma - \rho > 1 + \alpha$. No conditional mean limit for the curvature observation or interior condition on A is used.

For quantile regression in Corollary C.5, local density regularity identifies A_τ and controls the mean remainder. The mixed $(2+\nu)$ moment controls both tails and indicator crossings. The proof in Section D gives quadratic-mean exponent $\chi = \nu$. With bounded density envelopes, only $\mathbb{E}\|X\|^{2+\nu} < \infty$ is needed. Koenker (2005) gives standard local quantile conditions. Bounded regressors and moments above order three are not required.

The fixed-replica calibration results in Theorem 3.5 and Corollaries C.15 and C.16 use a deterministic conditional Gaussian limit and conditionally independent multiplier sequences. Applying Lemma C.26 to any fixed number of replicas gives a joint limit with independent Gaussian replica errors. This yields exact limiting Student or F calibration without estimating the conditional distribution. The argument is the cheap-bootstrap construction (Lam & Wang, 2026), combined here with the algorithm-specific paired linear representations. Neither $B \to \infty$ nor a consistent covariance estimator is needed. Scalar inference requires $B \geq 1$, and a q-dimensional ellipsoid requires $B \geq q$.

For covariance consistency and simultaneous output in Proposition C.18 and Corollary C.19, conditional uniform integrability controls empirical second-moment tails, which weak convergence alone does not control. The condition $B_n \to \infty$ makes the empirical approximation consistent. A sufficient condition for conditional uniform integrability is $\mathbb{E}^\star\|T_n^\star\|^{2+\epsilon} = O_p(1)$ for some $\epsilon > 0$.

Empirical plots of iterate decay, covariance stabilization, curvature increments, and bandwidth sensitivity are useful falsification checks, but they do not replace the probability conditions above. Section 4 reports coverage, coordinate bias, the paired frozen-joint distance, and finite-difference bias contrasts; it does not supply every diagnostic listed above.

B.2 Theory-level comparison

Cheap online bootstrap (Lam & Wang, 2026) uses a fixed number of online weighted averaged-SGD paths and Student calibration. Its zero-momentum, identity-preconditioner recursion with exponential weights and its scalar interval are exact special cases of the present framework. Lam and Wang also analyze an offline method, derive refined Berry-Esseen bounds, and give a two-stage sparse linear-regression procedure. The present extension covers momentum, nonsmooth scores, dynamic preconditioning, and several evaluations from one observation, but remains fixed-dimensional and asymptotic.

Online weighted SGD bootstrap (Fang et al., 2018) uses mean-one recursive multipliers and Polyak-Ruppert averages. Its recursion is the exact special case $\beta = 0$ and $\mathrm{P} = \mathrm{I}_d$ of the present framework. Its conventional empirical calibration corresponds to the increasing-replica output here. That method does not include momentum, dynamic preconditioning, or several evaluations from one observation.

The averaged fixed-momentum CLT (Tang et al., 2023) establishes first-order equivalence between averaged SGDM and averaged SGD. It provides a smooth original-path result with a step size that is constant within each run and delayed averaging. The present decreasing-step-size weighted representation gives conditional bootstrap inference and allows locally nonsmooth scores through quadratic-mean continuity.

The dynamic-preconditioner CLT (An & Huo, 2026a) provides the exact Abel decomposition and the threshold $\beta_{\mathrm{P}} > (1 + \alpha)/2$. We use this original-path threshold rather than claim it as a new result. The added results concern frozen and jointly weighted paths, paired equivalence, and the gain and curvature moment condition $2\gamma - \rho > 1 + \alpha$.

The SA-Adam Polyak-Ruppert CLT (An & Huo, 2026b) uses augmented-state analysis for momentum and non-convergent adaptive preconditioning. It treats a combined original algorithm with a sublinearly vanishing momentum gain. The present paper does not cover that combined recursion. Instead, it gives conditional multiplier inference for two separate classes: fixed momentum with fixed preconditioning, and dynamic preconditioning without momentum.

The linear-SA multiplier bootstrap (Samsonov et al., 2024) uses online perturbed replicas and covariance-free calibration. It gives nonasymptotic Gaussian and bootstrap error bounds for linear stochastic approximation. The present results are asymptotic, but cover nonlinear momentum recursions and recursions with a jointly evolving preconditioner.

The smooth nonlinear-SGD multiplier bootstrap (Sheshukova et al., 2026) provides a nonlinear Polyak-Ruppert bootstrap without explicit covariance estimation and gives nonasymptotic rates under smoothness. The present focus is state augmentation, nonsmooth local continuity, and dynamic secant stabilization.

Zeroth-order Polyak-Ruppert inference (Jin et al., 2021; Chen et al., 2024) combines random-direction finite differences with online uncertainty quantification. It derives asymptotic covariance and covariance-based confidence procedures for zeroth-order averages. The present weighting results for multiple evaluations from one observation instead characterize how evaluations from the same observation must be weighted inside a recursive multiplier bootstrap.

Online Newton with accelerated sketching (Wang et al., 2026) uses a dynamic second-order state for fully online inference. It targets the last iterate of a Hessian-averaged Newton method and estimates a Lyapunov covariance that includes noise from randomized computation. The present result concerns Polyak-Ruppert averages and a first-order law whose leading term does not contain the preconditioner in the stabilization regime.

Plug-in and covariance estimators (Chen et al., 2020; Zhu et al., 2023; Singh et al., 2025; Jiang et al., 2025; Wei et al., 2026) estimate the same sandwich covariance or its nonsmooth analogue. They include derivative-based, batch-means, equal-batch, nonsmooth, and de-biased Hessian-free constructions, without maintaining B perturbed optimization paths. They are methodological alternatives, but are not compared numerically in Section 4.

Random scaling and HiGrad (Lee et al., 2022; Su & Zhu, 2023) avoid direct sandwich estimation in online inference. They use self-normalization or branching and decorrelation rather than an empirical conditional law of weighted replicas. Their memory requirements, critical values, and finite-sample pivots differ from those of the present framework.

These comparisons concern the stated published or preprint theory and do not imply that one method uniformly dominates another. Computational performance depends on dimension, the number of replicas, conditioning, and the desired inferential output.

C THEORY SUPPLEMENT AND IMPLEMENTATION

C.1 METHODOLOGY

We construct online confidence procedures by evolving multiplier replicas alongside a stochastic approximation algorithm and comparing their Polyak-Ruppert averages with the original average. The construction applies a common random weight to all contributions from the same observation, thereby retaining the dependence induced by repeated evaluations of that observation. We develop this construction for fixed normalized momentum and for two forms of dynamic preconditioning. The resulting replica averages provide a common basis for inference.

C.1.1 ESTIMATING EQUATIONS AND MULTIPLIER REPLICAS

Let $d \geq 1$ be fixed, and let $\theta_\star \in \mathbb{R}^d$ be the unique local solution of $g(\theta) = 0$, where $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. At iteration $t \geq 1$, the algorithm receives a fresh block $\mathcal{U}_t$ containing a primary observation Z_t and any auxiliary observations, query directions, or sketches required for the update. Write $\mathcal{F}_t = \sigma(\mathcal{F}_0, \mathcal{U}_1, \ldots, \mathcal{U}_t)$, where $\mathcal{F}_0$ contains the original initialization and is independent of subsequent sampling blocks. We assume that $\mathcal{U}_t$ is independent of $\mathcal{F}_{t-1}$. The original algorithm and its replicas use this common block, but evaluate the required functions at their own parameter values. These evaluations therefore need not be independent of the preceding history.

A direct score $G(\theta, Z_t) \in \mathbb{R}^d$ satisfies $\mathbb{E}\{G(\theta, Z_t) \mid \mathcal{F}_{t-1}\} = g(\theta)$. To include approximate scores, let $\widehat{G}_t(\cdot)$ denote the random score map used at iteration t and suppose that, for each fixed θ in a neighborhood of $\theta_\star$,

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

where $b_t(\theta)$ is an $\mathcal{F}_{t-1}$-measurable approximation bias. Define the centered score error by $\xi_t(\theta) = \widehat{G}_t(\theta) - g(\theta) - b_t(\theta)$ and its value at the target by $\psi_t = \xi_t(\theta_\star)$. Then $\mathbb{E}(\psi_t \mid \mathcal{F}_{t-1}) = 0$. The direct-score formulation is recovered by taking $\widehat{G}_t(\theta) = G(\theta, Z_t)$ and $b_t \equiv 0$.

For a generic replica, let $(W_t)_{t\geq 1}$ be independent and identically distributed multipliers, independent of the original sampling stream and initialization, with

$$\mathbb{E}W_t = 1, \qquad \mathrm{Var}(W_t) = 1. \tag{12}$$

Thus $\mathbb{E}W_t^2 = 2$. One choice is $\mathbb{P}(W_t = 0) = \mathbb{P}(W_t = 2) = 1/2$. The construction does not require bounded or nonnegative multipliers, and no higher-order multiplier moment is imposed. Its validity nevertheless requires the conditional Lindeberg and weighted-recursion stability conditions stated in the corresponding results. For a fixed parameter, independence and the unit-mean condition give $\mathbb{E}\{W_t\widehat{G}_t(\theta) \mid \mathcal{F}_{t-1}\} = g(\theta) + b_t(\theta)$. Weighting therefore preserves the conditional mean of the score, including its approximation bias. Inference for $\theta_\star$ requires the contribution of this bias to be controlled as specified in those results.

We implement $B \geq 1$ replicas, with primary multipliers $W_t^{(b)}$ for $b = 1, \ldots, B$. Their multiplier sequences are independent across replicas, whereas their observations and query randomness are shared. Auxiliary observations receive weights according to the sampling-unit convention in Section 2.2. Let $\mathcal{F}_t^{(b)}$ be generated by $\mathcal{F}_t$ and all multipliers used by replica b through iteration t. The fresh block is independent of both $\mathcal{F}_{t-1}$ and $\mathcal{F}_{t-1}^{(b)}$, and the current multipliers are independent of the block and both histories. Conditional mean identities and score-moment bounds are understood with respect to the relevant pre-update history. In particular, bounds that hold uniformly in the parameter apply to the predictable parameter values of either the original recursion or a replica.

All parameter paths start from the same deterministic value $\theta_0 \in \mathbb{R}^d$ and use step sizes $\eta_t = \eta_0 t^{-\alpha}$, where $\eta_0 > 0$ and the fixed exponent α satisfies the conditions of the corresponding recursive result. Method-specific momentum and curvature states are initialized as described below. Writing $\theta_t^{(b)}$ generically for a replica parameter, define the Polyak–Ruppert averaging as

$$\bar{\theta}_n = \frac{1}{n}\sum_{t=1}^{n}\theta_t, \qquad \bar{\theta}_n^{(b)} = \frac{1}{n}\sum_{t=1}^{n}\theta_t^{(b)}. \tag{13}$$

For a result about a generic replica, a superscript $\star$ replaces its index (b); in comparisons, F and J identify the frozen and joint replicas. Thus $\theta_t^\star$, $m_t^\star$, $\mathrm{B}_t^\star$, $\mathrm{P}_t^\star$, and $\mathcal{F}_t^\star$ refer to the corresponding weighted path and its history. Write $\mathcal{H}_{t-1}$ for the relevant original or augmented pre-update history. These full averages are used throughout. A deterministic burn-in may instead omit the first k_n iterates and normalize by $n - k_n$, where k_n is an integer with $0 \le k_n < n$. Under the bounded-localization conditions of the recursive results, this change is first-order negligible when $k_n = o(\sqrt{n})$.

The forward conditional expectations used to analyze the updates are distinct from the conditional law used for bootstrap inference. At a reporting time n, let $\mathbb{P}^\star$ denote the regular conditional law of the replicas given $\mathcal{F}_n$, with the dependence on n suppressed. The original law is denoted by $\mathbb{P}$, and the joint law by $\mathbb{P} \otimes \mathbb{P}^\star$. We write $\mathbb{E}^\star$, $\mathrm{Cov}^\star$, and $\mathcal{L}^\star$ for expectation, covariance, and distribution under this conditional law. A term $R_n = o_{\mathbb{P}^\star}(1)$ in $\mathbb{P}$-probability means that $\mathbb{P}^\star(\|R_n\| > \varepsilon) \to 0$ in $\mathbb{P}$-probability for every $\varepsilon > 0$. Conditional on the original stream, replicas are driven by independent multiplier sequences.

C.1.2 Observation-level perturbations

We assign one random multiplier to each independent observation. All contributions computed from the same observation share this multiplier. This includes contributions from multiple function evaluations or different random query directions. Contributions from distinct independent observations receive independent multipliers. This rule also applies when those observations arrive in the same block $\mathcal{U}_t$. Write $W_t^{(b)}$ for the primary multiplier and $V_{tj}^{(b)}$ for the multiplier assigned to auxiliary observation j. Each has the law in (12), with independence across distinct sampling units, iterations, and replicas, and from the original random inputs.

This convention is consequential when a score is assembled from several function evaluations. For a scalar loss $L(\theta, Z_t)$, a deterministic radius $h_t > 0$, and a direction $U_t \in \mathbb{R}^d$, consider the central finite-difference score

$$D_{h_t}(\theta, Z_t, U_t) = \frac{L(\theta + h_t U_t, Z_t) - L(\theta - h_t U_t, Z_t)}{2h_t} U_t. \tag{14}$$

We apply the same multiplier W_t to both loss evaluations, yielding the weighted score $W_t D_{h_t}(\theta, Z_t, U_t)$.

Assigning independent multipliers W_t^+ and W_t^- to the two evaluations instead gives $\{W_t^+ L(\theta + h_t U_t, Z_t) - W_t^- L(\theta - h_t U_t, Z_t)\} U_t/(2h_t)$, which is not an observation-level perturbation of (14). When this finite-difference construction is used as $\widehat{G}_t$, its radius and direction scheme must also satisfy the score-moment and approximation-bias conditions of the relevant validity result.

The same reasoning applies to several query directions evaluated on one observation. For an integer $m \ge 1$, let $D_{t1}, \ldots, D_{tm} \in \mathbb{R}^d$ be the corresponding contributions at a given parameter value. Their weighted sum is $W_t \sum_{j=1}^m D_{tj}$. Assigning independent weights to these contributions generally changes their cross-covariances. This change may be negligible when the relevant cross-covariances vanish or the affected terms do not contribute to the first-order limiting distribution. Separate weights follow the observation-level resampling rule only when the contributions come from independent sampling units. The original estimator must also treat these units separately. Other weighting rules require their own theoretical justification. We use the observation-level rule for both parameter and curvature updates below.

C.1.3 Fixed normalized momentum

We first apply the construction to a direct-score recursion with fixed normalized momentum. Let $\beta \in [0, 1)$ and let $\mathrm{P} \in \mathbb{R}^{d \times d}$ be a fixed nonsingular preconditioning matrix. Starting from the initial θ_0 and $m_0 = 0 \in \mathbb{R}^d$, the original algorithm updates

$$m_t = \beta m_{t-1} + (1 - \beta) G(\theta_{t-1}, Z_t), \tag{15}$$

$$\theta_t = \theta_{t-1} - \eta_t \mathrm{P} m_t. \tag{16}$$

Replica b starts from $\theta_0^{(b)} = \theta_0$ and $m_0^{(b)} = m_0$ and applies the same momentum filter to the weighted score,

$$m_t^{(b)} = \beta m_{t-1}^{(b)} + (1 - \beta) W_t^{(b)} G(\theta_{t-1}^{(b)}, Z_t), \tag{17}$$

$$\theta_t^{(b)} = \theta_{t-1}^{(b)} - \eta_t \mathrm{P} m_t^{(b)}. \tag{18}$$

The original algorithm and each replica use the same observation Z_t. They evaluate their scores at θ_{t-1} and $\theta_{t-1}^{(b)}$, respectively. In replica b, the multiplier $W_t^{(b)}$ is applied to the current score before the momentum update. This weighted score enters the momentum state and affects subsequent parameter updates. Applying a multiplier to the accumulated momentum or the reported average does not reproduce this observation-level weighting.

We base inference on the averaged iterates defined in (13). The validity results require stability of both the original and replica recursions. These stability requirements are additional to the assumption that P is nonsingular.

C.1.4 Dynamic preconditioning

We next consider stochastic approximation with a time-varying preconditioner. In this setting, the parameter update uses the current score directly, without momentum. Dynamic preconditioning retains the observation-level weighting rule but introduces a choice about how the preconditioner is propagated through a replica. A frozen replica reuses the original preconditioner sequence, whereas a joint replica updates a separate curvature state. Both are based on the original recursion

$$\theta_t = \theta_{t-1} - \eta_t \mathrm{P}_t \widehat{G}_t(\theta_{t-1}), \qquad \mathrm{P}_t \in \mathbb{R}^{d\times d} \text{ is } \mathcal{F}_{t-1}\text{-measurable.} \tag{19}$$

The predictability requirement means that the preconditioner used at iteration t is determined before the current sampling block is observed. A common implementation obtains it by inverting a regularized symmetric curvature state. Fix $0 < \ell < u < \infty$, define $\mathcal{S}_{\ell,u} = \{\mathrm{M} \in \mathbb{R}^{d\times d} : \mathrm{M} = \mathrm{M}^{\mathsf{T}},\ \ell \mathrm{I}_d \preceq \mathrm{M} \preceq u\mathrm{I}_d\}$, and let $\Pi_{\ell,u}$ be the Frobenius projection onto this set. Here I_d is the identity matrix, $\preceq$ denotes positive-semidefinite ordering, and the projection clips the eigenvalues of a symmetric input to $[\ell, u]$.

Let $\widehat{\mathrm{H}}_t(\theta)$ be a symmetric curvature observation, such as a stochastic Hessian, score secant, sketched Hessian, or symmetrized random-direction estimate. With a deterministic learning rate $\lambda_t > 0$, initialize $\mathrm{B}_0 \in \mathcal{S}_{\ell,u}$ deterministically and set $\mathrm{P}_1 = \mathrm{B}_0^{-1}$. The curvature update is

$$\mathrm{B}_t = \Pi_{\ell,u}\left[\mathrm{B}_{t-1} + \lambda_t\{\widehat{\mathrm{H}}_t(\theta_{t-1}) - \mathrm{B}_{t-1}\}\right], \qquad \mathrm{P}_{t+1} = \mathrm{B}_t^{-1}. \tag{20}$$

Thus, the information collected at iteration t first affects the parameter update at iteration $t+1$, as required by (19). The sequence $\{\lambda_t\}$ is chosen to satisfy the stability and preconditioner-increment conditions of the applicable result. We use B_t for the curvature matrix and B for the number of replicas.

The **frozen-preconditioner** scheme introduces no additional curvature state. Starting from $\theta_0^{F(b)} = \theta_0$, replica b follows

$$\theta_t^{F(b)} = \theta_{t-1}^{F(b)} - \eta_t \mathrm{P}_t W_t^{(b)} \widehat{G}_t(\theta_{t-1}^{F(b)}). \tag{21}$$

The term frozen refers to sharing the realized original preconditioner sequence across replicas, not to keeping P_t constant over time. Each replica therefore requires only its parameter state and running average, while continuing to evaluate the score at its own parameter value.

The **joint** scheme instead propagates the observation-level perturbations through both the parameter and curvature states. Let $\widehat{\mathrm{H}}_t^{(b)}(\theta)$ denote the curvature observation formed at θ from the shared current sampling units with replica-specific observation-level multipliers. Initialize $\theta_0^{J(b)} = \theta_0$, $\mathrm{B}_0^{(b)} = \mathrm{B}_0$, and $\mathrm{P}_1^{(b)} = \mathrm{B}_0^{-1}$, and update

$$\theta_t^{J(b)} = \theta_{t-1}^{J(b)} - \eta_t \mathrm{P}_t^{(b)} W_t^{(b)} \widehat{G}_t(\theta_{t-1}^{J(b)}), \tag{22}$$

$$\mathrm{B}_t^{(b)} = \Pi_{\ell,u}\left[\mathrm{B}_{t-1}^{(b)} + \lambda_t\{\widehat{\mathrm{H}}_t^{(b)}(\theta_{t-1}^{J(b)}) - \mathrm{B}_{t-1}^{(b)}\}\right], \tag{23}$$

$$\mathrm{P}_{t+1}^{(b)} = (\mathrm{B}_t^{(b)})^{-1}.$$

Both inputs are evaluated at the pre-update parameter $\theta_{t-1}^{J(b)}$. In particular, $\mathrm{P}_t^{(b)} = (\mathrm{B}_{t-1}^{(b)})^{-1}$ is $\mathcal{F}_{t-1}^{(b)}$-measurable and does not depend on the current multipliers. If the primary observation also

contributes to curvature estimation, its multiplier enters $\mathrm{B}_t^{(b)}$ but affects the preconditioner only at the next iteration. This ordering preserves predictability while respecting the dependence between score and curvature contributions from a shared observation.

A score-secant construction makes the joint curvature perturbation explicit. Let $R_t \geq 1$ be a deterministic integer and $\delta_t > 0$ a deterministic perturbation radius. For an independent auxiliary batch, take $Z_{t1}^H, \ldots, Z_{tR_t}^H$ to have the same distribution as Z_t and to be independent of one another, Z_t, and $\mathcal{F}_{t-1}$. Draw directions $U_{tj} \in \mathbb{R}^d$ independently across j and independently of these observations and $\mathcal{F}_{t-1}$, with $\mathbb{E}(U_{tj}U_{tj}^{\mathsf{T}}) = \mathrm{I}_d$. For a square matrix, write $\mathrm{sym}(\mathrm{M}) = (\mathrm{M} + \mathrm{M}^{\mathsf{T}})/2$. Using the same observation and direction at the two secant endpoints, define

$$\mathrm{H}_{tj}(\theta) = \mathrm{sym}\left[\frac{G(\theta + \delta_t U_{tj}, Z_{tj}^H) - G(\theta - \delta_t U_{tj}, Z_{tj}^H)}{2\delta_t} U_{tj}^{\mathsf{T}}\right], \qquad \widehat{\mathrm{H}}_t(\theta) = \frac{1}{R_t}\sum_{j=1}^{R_t} \mathrm{H}_{tj}(\theta). \tag{24}$$

For replica b, the weighted curvature input is $\widehat{\mathrm{H}}_t^{(b)}(\theta) = R_t^{-1}\sum_{j=1}^{R_t} V_{tj}^{(b)}\mathrm{H}_{tj}(\theta)$. Each multiplier $V_{tj}^{(b)}$ is attached to one auxiliary observation. It applies to both score evaluations used to form $\mathrm{H}_{tj}(\theta)$. Independent auxiliary observations receive independent multipliers. When the auxiliary batch is independent of the primary observation, these multipliers are also independent of $W_t^{(b)}$. If an auxiliary contribution uses the primary observation, it receives the same multiplier as the primary contribution. For example, when $Z_{t1}^H = Z_t$, we set $V_{t1}^{(b)} = W_t^{(b)}$. The remaining independent observations retain independent multipliers.

When the argument is suppressed in subsequent moment bounds, a curvature input is evaluated at its own path's pre-update parameter. The original algorithm evaluates its curvature input at θ_{t-1}. The joint replica evaluates its weighted curvature input at $\theta_{t-1}^{J(b)}$. Thus, the joint replica recomputes the curvature contributions at its own parameter using the shared observations and directions. We do not impose unbiasedness or consistency of the curvature estimator as separate assumptions here. The theoretical results require stability of the parameter recursions and control of the changes in the preconditioners between successive iterations. Properties such as curvature unbiasedness or consistency may help verify these requirements.

C.1.5 Online inference and implementation

The three recursions differ in how algorithmic states are propagated, but lead to the same inferential quantities. For the selected method, use $\theta_t^{(b)}$ for its replica parameter and form the averages in (13). Define

$$\Delta_{n,b} = \bar{\theta}_n^{(b)} - \bar{\theta}_n, \qquad T_n^{(b)} = \sqrt{n}\,\Delta_{n,b}. \tag{25}$$

For a generic conditional replica, write $T_n = \sqrt{n}(\bar{\theta}_n^{\star} - \bar{\theta}_n)$. Fix a nominal error probability $\zeta \in (0,1)$. We consider two calibrations of these differences, one using the joint behavior of a fixed number of replicas and the other approximating their conditional distribution with an increasing number of replicas. Coverage in either case is subject to the assumptions of the corresponding validity result, including the required control of approximation bias.

When B is fixed, inference is based on the joint limit of the original estimation error and the B replica differences. For a fixed contrast $a \in \mathbb{R}^d$ with positive limiting variance, define $s_{n,B}^2(a) = B^{-1}\sum_{b=1}^{B}(a^{\mathsf{T}}\Delta_{n,b})^2$ and $s_{n,B}(a) = \{s_{n,B}^2(a)\}^{1/2}$. For $p \in (0,1)$, let $t_{B,p}$ be the p-quantile of a Student distribution with B degrees of freedom. The fixed-replica interval for $a^{\mathsf{T}}\theta_\star$ is

$$\left[a^{\mathsf{T}}\bar{\theta}_n - t_{B,1-\zeta/2}s_{n,B}(a), \quad a^{\mathsf{T}}\bar{\theta}_n + t_{B,1-\zeta/2}s_{n,B}(a)\right]. \tag{26}$$

Here $s_{n,B}^2(a)$ is a second moment about the original estimate, rather than a variance centered at the replica mean. Both its divisor and the Student degrees of freedom are B. This construction permits $B = 1$ for scalar inference and does not require an empirical approximation to the conditional bootstrap distribution.

The same fixed-replica calibration extends to a vector of contrasts. Let $1 \leq q \leq d$ and let $\mathrm{C} \in \mathbb{R}^{q\times d}$ be fixed with full row rank. Suppose that the projected estimation error has a nonsingular limiting covariance. Define $\mathrm{D}_{n,B}(\mathrm{C}) \in \mathbb{R}^{B\times q}$ to have bth row $(\mathrm{C}\Delta_{n,b})^{\mathsf{T}}$, set

$\widehat{\Sigma}_{n,B}(\mathrm{C}) = B^{-1}\mathrm{D}_{n,B}(\mathrm{C})^{\mathsf{T}}\mathrm{D}_{n,B}(\mathrm{C})$, and write $\mathcal{R}_{n,B}(\mathrm{C}) = \{\operatorname{rank}\mathrm{D}_{n,B}(\mathrm{C}) = q\}$. For $B \geq q$, let $c_{q,B}(\zeta) = \{qB/(B-q+1)\}F_{q,B-q+1}(1-\zeta)$, where $F_{r,s}(p)$ is the p-quantile of an F distribution with positive integer degrees of freedom r and s. The confidence region for $\mathrm{C}\theta_\star$ is

$$\mathcal{E}_{n,B}(\mathrm{C}) = \begin{cases} \left\{v \in \mathbb{R}^q : (\mathrm{C}\bar{\theta}_n - v)^{\mathsf{T}}\widehat{\Sigma}_{n,B}(\mathrm{C})^{-1}(\mathrm{C}\bar{\theta}_n - v) \leq c_{q,B}(\zeta)\right\}, & \mathcal{R}_{n,B}(\mathrm{C}), \\ \varnothing, & \mathcal{R}_{n,B}(\mathrm{C})^c. \end{cases} \tag{27}$$

The empty-set convention makes the procedure well defined when the replica differences do not span the contrast space and counts that event as noncoverage. On $\mathcal{R}_{n,B}(\mathrm{C})$, taking $q = 1$ and $\mathrm{C} = a^{\mathsf{T}}$ recovers the squared form of (26), since $F_{1,B}(1-\zeta) = t^2_{B,1-\zeta/2}$.

An alternative is to approximate the conditional law of T_n using $B = B_n \to \infty$ replicas. For a fixed contrast $a \in \mathbb{R}^d$ and $p \in (0,1)$, define the left conditional quantile by $q_{p,n}(a) = \inf\{x \in \mathbb{R} : \mathbb{P}^\star(a^{\mathsf{T}}T_n \leq x) \geq p\}$. The corresponding ideal basic interval is

$$\left[a^{\mathsf{T}}\bar{\theta}_n - n^{-1/2}q_{1-\zeta/2,n}(a), \quad a^{\mathsf{T}}\bar{\theta}_n - n^{-1/2}q_{\zeta/2,n}(a)\right]. \tag{28}$$

Let $\widehat{q}_{p,n,B}(a)$ be the empirical p-quantile of $\{a^{\mathsf{T}}T_n^{(b)} : 1 \leq b \leq B\}$ under the same left-quantile convention. Replacing the conditional quantiles by these empirical quantiles gives the implemented interval

$$\left[a^{\mathsf{T}}\bar{\theta}_n - n^{-1/2}\widehat{q}_{1-\zeta/2,n,B}(a), \quad a^{\mathsf{T}}\bar{\theta}_n - n^{-1/2}\widehat{q}_{\zeta/2,n,B}(a)\right]. \tag{29}$$

For $B \geq 2$, the same replicas provide a covariance estimate for the root-n error. Write $\bar{T}_{n,B} = B^{-1}\sum_{b=1}^{B} T_n^{(b)}$ for their mean and define

$$\widehat{\mathrm{V}}_{n,B} = \frac{1}{B-1}\sum_{b=1}^{B}(T_n^{(b)} - \bar{T}_{n,B})(T_n^{(b)} - \bar{T}_{n,B})^{\mathsf{T}}. \tag{30}$$

Under the conditional uniform-integrability condition in Proposition C.18, this matrix consistently estimates the limiting covariance of T_n as $n, B_n \to \infty$, whereas $\widehat{\mathrm{V}}_{n,B}/n$ is the sample covariance of the unscaled differences $\Delta_{n,b}$. Unlike $\widehat{\Sigma}_{n,B}(\mathrm{C})$ in fixed-replica inference, it is centered at the replica mean and uses the divisor $B-1$.

The covariance estimate also provides a common scale for simultaneous inference on a finite collection of fixed contrasts $a_1, \ldots, a_q \in \mathbb{R}^d$. Here $q \geq 1$ counts the contrasts and may exceed d. When $a_j^{\mathsf{T}}\widehat{\mathrm{V}}_{n,B}a_j > 0$ for every j, compute

$$M_{n,b} = \max_{1\leq j\leq q}\frac{|a_j^{\mathsf{T}}T_n^{(b)}|}{(a_j^{\mathsf{T}}\widehat{\mathrm{V}}_{n,B}a_j)^{1/2}}, \qquad b = 1, \ldots, B. \tag{31}$$

Let $\widehat{c}_{1-\zeta,n,B}$ be the empirical $(1-\zeta)$-quantile of $M_{n,1}, \ldots, M_{n,B}$. The simultaneous intervals are

$$a_j^{\mathsf{T}}\bar{\theta}_n \pm \frac{\widehat{c}_{1-\zeta,n,B}}{\sqrt{n}}(a_j^{\mathsf{T}}\widehat{\mathrm{V}}_{n,B}a_j)^{1/2}, \qquad j = 1, \ldots, q. \tag{32}$$

If any estimated contrast variance is nonpositive, the simultaneous output is defined to be empty. Under the covariance-consistency conditions of the corresponding validity result and positive limiting contrast variances, the probability of this event tends to zero.

Empirical quantiles, covariance estimates, and simultaneous critical values are used here in the increasing-replica regime. For each terminal sample size n, all B_n replicas must be maintained from the beginning of that run. The specified algorithm does not introduce new replicas partway through the stream. All coverage guarantees concern a fixed reporting time, so reporting at several times does not in itself provide simultaneous coverage over time.

Algorithm 1 implements the recursions and their inferential outputs in one pass. At each iteration, observations and query randomness are shared, but every score and joint curvature input is evaluated at the pre-update parameter of the path that uses it. The current block is discarded only after all such evaluations are complete.

Algorithm 1 Online multiplier inference for Polyak-Ruppert averages

Require: Common initial parameter θ_0, integer replica count $B \geq 1$, step sizes η_t, multiplier law (12), and method in {momentum, frozen, joint}.

Require: Method-specific initial states and tuning sequences, predetermined reporting times, error probability ζ, contrasts, and output in {fixed-replica, empirical}.

Initialize all original and replica states for the selected method. Set $\bar{\theta}_0 = 0$ and $\bar{\theta}_0^{(b)} = 0$ for $b = 1, \ldots, B$.
for $t = 1, 2, \ldots$ **do**
  Receive $\mathcal{U}_t$ and identify its independent observation-level units.
  if method is momentum **then**
    Evaluate $G(\theta_{t-1}, Z_t)$ and update the original momentum and parameter by (15) to (16).
  else
    Evaluate $\widehat{G}_t(\theta_{t-1})$ and update θ_t by (19) using P_t.
    For the projected-curvature construction, evaluate $\widehat{\mathrm{H}}_t(\theta_{t-1})$ and update B_t and P_{t+1} by (20). Retain P_t for the current frozen-replica updates.
  for $b = 1, \ldots, B$ **do**
    Draw the primary multiplier $W_t^{(b)}$.
    if method is momentum **then**
      Evaluate $G(\theta_{t-1}^{(b)}, Z_t)$ and update the replica by (17) to (18).
    else if method is frozen **then**
      Evaluate $\widehat{G}_t(\theta_{t-1}^{F(b)})$ and update the replica by (21) using the retained original P_t.
    else
      Evaluate $\widehat{G}_t(\theta_{t-1}^{J(b)})$ and update the parameter by (22) using $\mathrm{P}_t^{(b)}$.
      Assign one multiplier to each independent auxiliary observation, reusing $W_t^{(b)}$ whenever the primary observation is reused.
      Form $\widehat{\mathrm{H}}_t^{(b)}(\theta_{t-1}^{J(b)})$ and update $\mathrm{B}_t^{(b)}$ and $\mathrm{P}_{t+1}^{(b)}$ by (23).
  Update $\bar{\theta}_t = \bar{\theta}_{t-1} + t^{-1}(\theta_t - \bar{\theta}_{t-1})$ and, for each b, $\bar{\theta}_t^{(b)} = \bar{\theta}_{t-1}^{(b)} + t^{-1}(\theta_t^{(b)} - \bar{\theta}_{t-1}^{(b)})$.
  Discard the current observations after all original and replica evaluations are complete.
  if t is a reporting time **then**
    Form $\Delta_{t,b}$ for $b = 1, \ldots, B$ using (25).
    if output is fixed-replica **then**
      Report (26) or (27), with the stated dimensional and rank requirements.
    else
      Form $T_t^{(b)} = \sqrt{t}\,\Delta_{t,b}$ and report (29), (30), or (32), with the stated variance requirements.

For a fixed number of replicas, the persistent state of this implementation does not grow with the number of observations. It does grow with B_n when the number of replicas increases. The following costs refer to updates and retained state, excluding storage of the current sampling block and additional reporting calculations. Let $c_G(d)$ denote the cost of one primary score evaluation. With a dense P, fixed normalized momentum requires $O\{(B+1)c_G(d) + (B+1)d^2\}$ time per iteration and $O\{(B+1)d + d^2\}$ state memory. Identity, diagonal, or $O(d)$-nonzero preconditioners reduce the matrix-vector cost to $O\{(B+1)d\}$. For a general sparse P, this cost is $O\{(B+1)\,\mathrm{nnz}(\mathrm{P})\}$, where $\mathrm{nnz}(\mathrm{P})$ is the number of nonzero entries.

The additional cost of dynamic preconditioning depends on whether curvature states are shared or replicated. A dense frozen implementation stores $O(d^2 + Bd)$ state and performs the original curvature construction, projection, and inversion once per iteration, in addition to primary score evaluations and replica matrix-vector products. Dense projection or inversion can cost $O(d^3)$. A dense joint implementation maintains $B+1$ curvature matrices, including the original one, requiring $O\{(B+1)d^2\}$ state memory. Its path-specific projections or inversions can cost $O\{(B+1)d^3\}$ per iteration, in addition to score and curvature evaluations. Under (24), each curvature path requires $2R_t$ auxiliary score evaluations per iteration. Structured curvature representations can reduce these costs, provided that the resulting updates satisfy the required stability and preconditioner-increment conditions.

C.2 THEORETICAL RESULTS

C.2.1 A CONDITIONAL TRANSFER THEOREM

Let d_{BL} be the bounded-Lipschitz distance on probability laws over $\mathbb{R}^d$:

$$d_{\mathrm{BL}}(\mu,\nu) = \sup_{\|f\|_\infty \le 1,\, \mathrm{Lip}(f)\le 1} \left| \int f\, d\mu - \int f\, d\nu \right|.$$

The influence vectors below may form a triangular array, which permits time-varying finite-difference scales, query distributions, or auxiliary batch sizes.

The assumptions separate distributional approximation from optimization. The transfer theorem takes paired linear representations as inputs. The momentum and dynamic results then prove those representations from score regularity, rates for both parameter paths, and the relevant control of algorithmic states. Primitive drift conditions provide sufficient routes to the rates; outside those classes, the rate and localization hypotheses must be checked for the chosen recursion.

Assumption C.1. *The multipliers satisfy* (12) *and are independent of the original data. There are observed vectors $\psi_{n1},\dots,\psi_{nn} \in \mathbb{R}^d$ and a positive-semidefinite matrix* Q *such that*

$$\frac{1}{n}\sum_{t=1}^{n} \psi_{nt}\psi_{nt}^{\mathsf{T}} \xrightarrow{p} \mathrm{Q}, \tag{33}$$

$$\mathcal{L}_n(\varepsilon) := \frac{1}{n}\sum_{t=1}^{n} \|\psi_{nt}\|^2 \mathbb{E}\left[(W_t-1)^2 \mathbf{1}\{|W_t-1|\,\|\psi_{nt}\| > \varepsilon\sqrt{n}\}\right] \xrightarrow{p} 0 \quad \textit{for every } \varepsilon>0, \tag{34}$$

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \psi_{nt} \Longrightarrow N(0,\mathrm{Q}). \tag{35}$$

The expectation in (34) *is over an independent copy of the multiplier, with the observed influence vectors held fixed.*

Assumption C.2. *This is a representation condition for the transfer step. It is established independently by Theorems 3.2 and 3.3 for the respective recursions. For a nonsingular matrix* A*,*

$$\sqrt{n}(\bar\theta_n - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_{nt} + r_n, \qquad r_n = o_p(1), \tag{36}$$

$$\sqrt{n}(\bar\theta_n^\star - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} W_t\psi_{nt} + r_n^\star, \qquad r_n^\star = o_{\mathbb{P}\otimes\mathbb{P}^\star}(1). \tag{37}$$

The main statement is Theorem 3.1.

The conditional multiplier central limit theorem used in Theorem 3.1 is classical (see Præstgaard & Wellner, 1993; van der Vaart & Wellner, 1996). The theorem is included to separate the conditional transfer step from the algorithm-specific linearization. Condition (34) is the ordinary Lindeberg condition for the observed triangular array. It is weaker than a $2+\delta$ moment assumption. In particular, if the influences are i.i.d. with only a finite second moment, then Assumption C.1 follows from the law of large numbers, the multivariate central limit theorem, and dominated convergence; see Lemma C.27. For martingale-difference arrays, a sufficient route is convergence of the averaged predictable covariance to Q together with a uniform $2+\delta$ moment for some $\delta>0$. A martingale moment inequality applied to centered coordinate products then gives empirical covariance convergence, and the moment bound gives the martingale Lindeberg condition and $\max_{t\le n}\|\psi_{nt}\|/\sqrt{n} \xrightarrow{p} 0$ (see Hall & Heyde, 1980). The latter maximum bound and tight empirical second moments also verify multiplier Lindeberg with only a finite second multiplier moment. A useful diagnostic is the ratio of the largest estimated influence norm to the square root of the total empirical second moment. Such a diagnostic can reveal domination by a few observations, but it does not replace (34).

C.2.2 FIXED-MOMENTUM BOOTSTRAP

Write $g(\theta) = \mathbb{E}G(\theta, Z_t)$, $\xi_t(\theta) = G(\theta, Z_t) - g(\theta)$, and $\psi_t = G(\theta_\star, Z_t)$.

Assumption C.3. *The following conditions hold.*

(M1) ***Step size and local mean.*** *$\eta_t = \eta_0 t^{-\alpha}$ with $1/2 < \alpha < 1$; $\beta \in [0, 1)$ is fixed; and P is fixed and nonsingular. For some neighborhood $\mathcal{N}$ of $\theta_\star$, some nonsingular A, some $\nu \in (0, 1]$, and a constant C,*

$$g(\theta_\star + e) = \mathrm{A}e + r(e), \qquad \|r(e)\| \le C\|e\|^{1+\nu}, \qquad \alpha(1+\nu) > 1, \tag{38}$$

whenever $\theta_\star + e \in \mathcal{N}$.

(M2) ***Second moments and quadratic-mean continuity.*** *The target innovations satisfy Assumption C.1 with $\psi_{nt} = \psi_t$. For some $\chi \in (0, 2]$,*

$$\mathbb{E}\|\psi_t\|^2 < \infty, \qquad \mathbb{E}\|\xi_t(\theta) - \xi_t(\theta_\star)\|^2 \le C\|\theta - \theta_\star\|^\chi, \quad \theta \in \mathcal{N}.$$

(M3) ***Original and weighted stability.*** *Under the original law and the joint law, respectively,*

$$\mathbb{E}\|\theta_t - \theta_\star\|^2 \le C\eta_t, \qquad \mathbb{E}\|\theta_t^\star - \theta_\star\|^2 \le C\eta_t. \tag{39}$$

The initial parameter and momentum states have finite second moments under the corresponding laws.

(M4) ***Localization.*** *The preceding conditions may hold globally. Otherwise, for each n there are auxiliary original and weighted recursions governed by the same equations* (15) *to* (18)*, using globally defined measurable extensions that agree with the original score and mean field on $\mathcal{N}$ and satisfy (M1) to (M3) along the auxiliary paths. For a deterministic $k_n = o(\sqrt{n})$, each auxiliary path and its corresponding original path coincide for every $t \in [k_n, n]$ with probability tending to one under the corresponding law. The original and auxiliary states over the first k_n steps have uniformly bounded second moments. All bounds are uniform in the localization index n, and the extensions use the same target innovations and sampling blocks.*

The rate and localization requirements in (M3) and (M4) are model-specific in general. Convergence into $\mathcal{N}$ alone does not give (M4): an auxiliary path changed earlier need not subsequently agree exactly with the original path. The agreement and initial-block bounds are part of the hypothesis. The next result gives one primitive closure for a globally stable smooth mean field. It applies to the original and weighted paths under the same conditions.

Proposition C.4. *Let $\eta_t = \eta_0 t^{-\alpha}$ with $0 < \alpha < 1$, let $\beta \in [0, 1)$ be fixed, and let P be fixed and nonsingular. Suppose $g(\theta_\star) = 0$, g is globally Lipschitz, and there are a symmetric positive-definite matrix R and a constant $\mu > 0$ such that*

$$e^\mathsf{T}\mathrm{R}\mathrm{P}g(\theta_\star + e) \ge \mu\|e\|_\mathrm{R}^2, \qquad e \in \mathbb{R}^d, \tag{40}$$

where $\|e\|_\mathrm{R}^2 = e^\mathsf{T}\mathrm{R}e$. For every parameter error e predictable with respect to the relevant pre-update history, assume

$$\mathbb{E}\{G(\theta_\star + e, Z_t) \mid \mathcal{H}_{t-1}\} = g(\theta_\star + e), \tag{41}$$

$$\mathbb{E}\{\|G(\theta_\star + e, Z_t)\|^2 \mid \mathcal{H}_{t-1}\} \le C(1 + \|e\|^2) \tag{42}$$

almost surely. On a weighted path, the current multiplier is independent of the current sampling block and the pre-update history, with $\mathbb{E}W_t = 1$ and $\mathbb{E}W_t^2 < \infty$. If the initial parameter and momentum states have finite second moments, then

$$\mathbb{E}\|\theta_t - \theta_\star\|^2 + \mathbb{E}\|\theta_t^\star - \theta_\star\|^2 \le C'\eta_t, \qquad t \ge 1, \tag{43}$$

$$\sup_{t \ge 0}\{\mathbb{E}\|m_t\|^2 + \mathbb{E}\|m_t^\star\|^2\} < \infty. \tag{44}$$

Thus (M3) holds. If the remaining mean and score conditions also hold globally, the global branch of (M4) applies with $\mathcal{N} = \mathbb{R}^d$ and $k_n = 0$.

For a linear mean $g(\theta_\star + e) = \mathrm{A}e$, a concrete sufficient case is that PA has eigenvalues with positive real parts and the noise is additive, centered, and has a finite second moment. The positive-definite solution of $(\mathrm{PA})^\mathsf{T}\mathrm{R} + \mathrm{R}(\mathrm{PA}) = \mathrm{I}_d$ verifies the drift. For a globally strongly convex risk

with a globally Lipschitz gradient and positive-definite P, take $\mathrm{R} = \mathrm{P}^{-1}$ and verify the oracle moment bound. These are extensions of the familiar Lyapunov route to ordinary-SGD mean-square rates (Moulines & Bach, 2011); the momentum and weighted-path conclusions follow from the proposition above.

The main statement is Theorem 3.2.

The cancellation in (4) follows by summing the normalized momentum update: the score sum equals a step-size-weighted sum of parameter increments plus a momentum endpoint. The parameter rates make both terms negligible after division by $\sqrt{n}$, leaving the target innovation transformed by A^{-1}. Tang et al. (2023) provide a smooth original-path averaging precedent with a step size fixed within each run, batch-size conditions, and an initial averaging delay. Their asymptotic regime differs from the decreasing-step-size full average considered here. The result needed for online resampling is the weighted representation (5); the proof also permits nonsmooth sample scores through quadratic-mean continuity. The condition $\alpha(1+\nu) > 1$ is exactly what makes the accumulated nonlinear mean remainder negligible at the root-n scale. It is automatic for every $\alpha > 1/2$ when the derivative of g is locally Lipschitz. Only second moments are used for the target score and multiplier. A global primitive yielding bounded momentum second moments is given in Proposition C.4; the general asymptotic-linearization proof instead controls the momentum endpoint directly from the parameter rate.

The $O(\eta_t)$ bounds are optimization conditions. Proposition C.4 verifies them under a global Lyapunov drift; local models require their own stability and localization argument (Kushner & Yin, 2003; Moulines & Bach, 2011; Tang et al., 2023). Positive definiteness of P is a common way to obtain a stable drift when A is positive definite, but the inferential cancellation uses only that P and A are nonsingular. If an indefinite P is used, the stability assumption must be verified directly. Bounded nonnegative multipliers may facilitate that verification, but finite variance is sufficient once the weighted path rate is established. The result is pointwise in each fixed $\beta < 1$ and fixed nonsingular P: no uniform approximation as $\beta \uparrow 1$, or over sequences of increasingly ill-conditioned preconditioners, is asserted.

When $\beta = 0$ and $\mathrm{P} = \mathrm{I}_d$, (15) to (18) reduce exactly to the original and weighted SGD recursions (Fang et al., 2018).

Corollary C.5. *Let*

$$G_\tau(\theta; X, Y) = X\{\mathbf{1}(Y \le X^\mathsf{T}\theta) - \tau\}, \qquad \tau \in (0,1), \tag{45}$$

where (X_t, Y_t) are i.i.d. copies of (X, Y), and let θ_τ solve $\mathbb{E}G_\tau(\theta_\tau; X, Y) = 0$. Write $q_X = X^\mathsf{T}\theta_\tau$. Suppose:

(i) for some $r_0 > 0$ and nonnegative random envelopes $F(X)$ and $L(X)$,

$$\sup_{|v| \le r_0} f_{Y|X}(q_X + v \mid X) \le F(X),$$
$$|f_{Y|X}(q_X + v \mid X) - f_{Y|X}(q_X \mid X)| \le L(X)|v|^\nu, \qquad |v| \le r_0, \tag{46}$$

almost surely, for some $\nu \in (0,1]$;

(ii) the mixed moment matched to the density exponent satisfies

$$\mathbb{E}\left[(1 + F(X) + L(X))\|X\|^{2+\nu}\right] < \infty; \tag{47}$$

(iii)

$$\mathrm{A}_\tau = \mathbb{E}\{f_{Y|X}(X^\mathsf{T}\theta_\tau \mid X) XX^\mathsf{T}\} \tag{48}$$

is positive definite;

(iv) the recursions use $\eta_t = \eta_0 t^{-\alpha}$ with $1/2 < \alpha < 1$, a fixed $\beta \in [0,1)$, and a fixed nonsingular matrix P, and satisfy the original and weighted parameter-rate bounds (M3), including its initial-moment conditions, with $\alpha(1+\nu) > 1$.

Then (6) *holds with*

$$\mathrm{A} = \mathrm{A}_\tau, \qquad \mathrm{Q} = \mathrm{S}_\tau = \mathbb{E}\left[XX^\mathsf{T}\{\mathbf{1}(Y \le X^\mathsf{T}\theta_\tau) - \tau\}^2\right]. \tag{49}$$

The indicator changes only when the response lies between two fitted quantiles. Splitting according to whether $|X^{\mathsf{T}}e| \le r_0$ bounds local crossings by the density envelope and controls the complementary event by the mixed design moment. The proof gives the mean and score bounds for every $e \in \mathbb{R}^d$, so no auxiliary localization is needed here. Condition (iv) remains an independent path-rate requirement; conditions (i) to (iii) do not imply it. In particular, when $\mathbb{E}\|X\| < \infty$, the mean quantile score is bounded as a function of θ and therefore cannot satisfy the global quadratic drift in (40) over all of $\mathbb{R}^d$. The rate for the stated unprojected recursion must therefore be established by a separate model-specific argument. The remaining conditions are local in the conditional density and require no bounded regressors. The mixed $(2+\nu)$ moment controls the local mean remainder and the probability that an indicator crosses the random quantile interval. It is sufficient for quadratic-mean continuity with exponent $\chi = \nu$; no bounded-regressor condition is used. The target influence itself needs only a second moment. These conditions are consistent with standard local quantile-regression regularity (see Koenker, 2005). When the density envelopes are bounded, (47) reduces to $\mathbb{E}\|X\|^{2+\nu} < \infty$. A locally Lipschitz conditional density has $\nu = 1$ and therefore requires a third moment, not a moment of order greater than three. For smaller ν, the step-size condition $\alpha(1+\nu) > 1$ records the tradeoff between density smoothness and the decay of the nonlinear drift remainder. The case $\beta = 0$ is the quantile specialization (Fang et al., 2018).

C.2.3 Stabilizing dynamic preconditioning

The next conditions are stated for one original path and one generic weighted path. The weighted path can be frozen or joint. Write $e_t = \theta_t - \theta_\star$, $e_t^\star = \theta_t^\star - \theta_\star$, and

$$\mathrm{M}_t = (\mathrm{P}_t\mathrm{A})^{-1}, \qquad \mathrm{M}_t^\star = (\mathrm{P}_t^\star\mathrm{A})^{-1}.$$

Assumption C.6. *The following conditions hold.*

(D1) $\eta_t = \eta_0 t^{-\alpha}$ *with* $1/2 < \alpha < 1$. *The mean field satisfies* (38) *on a neighborhood* $\mathcal{N}$, *with* A *nonsingular,* $\nu \in (0,1]$, *and* $\alpha(1+\nu) > 1$.

(D2) The centered score error $\xi_t(\theta)$ *defined in Section 2.1 obeys, for some* $\chi \in (0,2]$,

$$\mathbb{E}\{\|\xi_t(\theta) - \xi_t(\theta_\star)\|^2 \mid \mathcal{H}_{t-1}\} \le C\|\theta - \theta_\star\|^\chi \tag{50}$$

for $\theta \in \mathcal{N}$ *predictable under the corresponding original or augmented history* $\mathcal{H}_{t-1}$, *and the target innovations satisfy Assumption C.1 with* $\psi_{nt} = \psi_t$.

(D3) Put $\Delta b_t(\theta) = b_t(\theta) - b_t(\theta_\star)$. *The target bias satisfies*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^n b_t(\theta_\star) \xrightarrow{p} 0, \qquad \frac{1}{n}\sum_{t=1}^n \|b_t(\theta_\star)\|^2 \xrightarrow{p} 0. \tag{51}$$

For each of the original and weighted paths, denoted generically by ϑ_t *and interpreted under its corresponding law,*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^n \mathbb{E}\|\Delta b_t(\vartheta_{t-1})\| \longrightarrow 0. \tag{52}$$

(D4) The original and weighted parameter paths satisfy

$$\mathbb{E}\|e_t\|^2 \le C\eta_t, \qquad \mathbb{E}\|e_t^\star\|^2 \le C\eta_t, \tag{53}$$

under their respective laws, and their initial parameter states have finite second moments. These conditions may hold globally. Otherwise, for each n *there are auxiliary original and weighted recursions governed by the same equations* (19), (21), *or* (22), *as applicable, using globally defined measurable extensions that agree with the original oracle and mean field on* $\mathcal{N}$ *and satisfy (D1) to (D3), the displayed parameter-rate bound, and Assumption C.7 for their own preconditioner sequences, under the corresponding laws and uniformly in* n. *The extensions preserve the target innovation and target bias. For a deterministic* $k_n = o(\sqrt{n})$, *each auxiliary path and its corresponding original path coincide for every* $t \in [k_n, n]$ *with probability tending to one under the corresponding law. The original and auxiliary states over the first* k_n *steps have uniformly bounded second moments. All bounds are uniform in the localization index* n, *and the extensions use the same target innovations and sampling blocks.*

A convenient sufficient condition for (52), with a deterministic nonnegative envelope $L_{b,t}$, is

$$\|\Delta b_t(\theta)\| \le L_{b,t}\|\theta - \theta_\star\|, \qquad \frac{1}{\sqrt{n}}\sum_{t=1}^{n} L_{b,t}\eta_t^{1/2} \to 0.$$

The theorem is stated with the direct pathwise conditions because they also cover approximation biases that are not uniformly Lipschitz in a neighborhood.

Assumption C.7. *The original and weighted preconditioners are predictable, and* $\mathbb{E}\|\mathrm{M}_1\|_{\mathrm{F}}^2 + \mathbb{E}\|\mathrm{M}_1^\star\|_{\mathrm{F}}^2 < \infty$. *Moreover,*

$$\Delta_{M,n} := \frac{1}{\sqrt{n}}\sum_{t=1}^{n-1} t^{\alpha/2}\{\mathbb{E}\|\mathrm{M}_{t+1} - \mathrm{M}_t\|_{\mathrm{F}}^2\}^{1/2} \longrightarrow 0, \tag{54}$$

$$\Delta_{M,n}^\star := \frac{1}{\sqrt{n}}\sum_{t=1}^{n-1} t^{\alpha/2}\{\mathbb{E}\|\mathrm{M}_{t+1}^\star - \mathrm{M}_t^\star\|_{\mathrm{F}}^2\}^{1/2} \longrightarrow 0. \tag{55}$$

If, for some $\beta_{\mathrm{P}} > 0$,

$$\{\mathbb{E}\|\mathrm{M}_{t+1} - \mathrm{M}_t\|_{\mathrm{F}}^2\}^{1/2} + \{\mathbb{E}\|\mathrm{M}_{t+1}^\star - \mathrm{M}_t^\star\|_{\mathrm{F}}^2\}^{1/2} \le Ct^{-\beta_{\mathrm{P}}},$$

then Assumption C.7 holds whenever

$$\beta_{\mathrm{P}} > \frac{1+\alpha}{2}. \tag{56}$$

An & Huo (2026a) show that the rate hypotheses alone cannot lower this threshold, using a construction in which the matrix and error sequences are prescribed separately. This does not establish necessity for every coupled adaptive recursion. The weighted variation condition itself is more general and permits non-polynomial stabilization.

The main statement is Theorem 3.3.

The original, unweighted part of Theorem 3.3 follows the preconditioner-isolation argument (An & Huo, 2026a); the weighted part is the bootstrap extension. The condition in Assumption C.7 is the quantity that appears directly in the Abel remainder

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \eta_t^{-1}\mathrm{M}_t^\star(e_{t-1}^\star - e_t^\star).$$

The polynomial threshold (56) is sufficient because it makes the matrix-increment part of this sum vanish. It is not needed as a separate structural assumption. The threshold is sharp for the rate hypotheses tested by the decoupled construction (An & Huo, 2026a), but it is not a pointwise necessity statement for every data-generating mechanism: special models can exhibit additional cancellation at or below the boundary. The leading influence is independent of the preconditioner because $\mathrm{M}_t^\star\mathrm{P}_t^\star = \mathrm{A}^{-1}$ exactly.

The assumptions distinguish optimization stability from inferential stabilization. Spectral clipping controls invertibility but does not by itself imply (53). Once the parameter path has the usual $O(\eta_t)$ rate and the effective inverse drift satisfies (54) to (55), the dynamic preconditioner is absent from the first-order average. For $\mathrm{P}_t^\star = (\mathrm{B}_{t-1}^\star)^{-1}$, the increment of $\mathrm{M}_t^\star = \mathrm{A}^{-1}\mathrm{B}_{t-1}^\star$ is proportional, up to fixed matrix norms, to the increment of the projected curvature state.

Corollary C.8. *Suppose* $\mathbb{E}\|\mathrm{M}_1\|_{\mathrm{F}}^2 < \infty$*, the original path satisfies* (54)*, and the frozen weighted parameter recursion* (21) *satisfies Assumption C.6. Then* (8) *and conditional bootstrap consistency hold with* $\mathrm{P}_t^\star = \mathrm{P}_t$.

Freezing is computationally attractive because no replica-specific curvature state is stored. Its validity is not an automatic consequence of using a consistent preconditioner; it follows from the stronger fact that the entire dynamic-preconditioner contribution is $o_p(n^{-1/2})$ in the average. At or below the stabilization boundary, a frozen bootstrap can omit a first-order random component and is not covered by this corollary.

Proposition C.9. *Consider the projected update* (23) *with* $\mathrm{B}_0^\star \in \mathcal{S}_{\ell,u}$. *If*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n-1} t^{\alpha/2} \lambda_t \left[1 + \{\mathbb{E}\|\widehat{\mathrm{H}}_t^\star\|_{\mathrm{F}}^2\}^{1/2} \right] \longrightarrow 0, \tag{57}$$

then (55) *holds. In particular, if*

$$0 < \lambda_t \le C t^{-\gamma}, \qquad \mathbb{E}\|\widehat{\mathrm{H}}_t^\star\|_{\mathrm{F}}^2 \le C t^{\rho}, \qquad \rho \ge 0, \tag{58}$$

then

$$\{\mathbb{E}\|\mathrm{M}_{t+1}^\star - \mathrm{M}_t^\star\|_{\mathrm{F}}^2\}^{1/2} \le C t^{-\gamma+\rho/2}, \tag{59}$$

and (57) *holds whenever*

$$2\gamma - \rho > 1 + \alpha. \tag{60}$$

These conclusions require neither $\mathbb{E}(\widehat{\mathrm{H}}_t^\star \mid \mathcal{F}_{t-1}^\star) \to \mathrm{A}$ *nor* $\mathrm{A} \in \mathrm{int}(\mathcal{S}_{\ell,u})$.

The removal of curvature consistency is substantive but conditional: it weakens the inferential assumptions, not the requirements for a useful optimization path. A consistent secant state can help verify (53) and improve finite-sample conditioning. It is not needed in the algebra that removes $\mathrm{P}_t^\star$ from the leading influence function. This distinction prevents an assumption used to assess optimization quality from being presented as a first-order bootstrap necessity.

The main statement is Theorem 3.4.

The two bootstrap schemes have the same first-order conditional law under the stated stabilization conditions but need not behave identically at moderate sample sizes. The joint scheme reproduces curvature-state transients, while the frozen scheme uses less memory and avoids replica-specific curvature noise. The theorem does not order their finite-sample coverage errors. Equation (10) gives a direct empirical check: the root-n distance between paired frozen and joint replica averages should shrink when first-order stabilization is adequate.

Corollary C.10. *Assume the parameter-rate and initial-moment bounds in (D4) for the original and joint paths. The bias bounds below hold at their pre-update parameter values almost surely. If the local branch of (D4) is used, impose these bounds and rates also along its auxiliary paths, uniformly in the localization index, and retain the exact-coupling and initial-block conditions. Suppose* $\lambda_t \asymp t^{-\gamma}$ *and*

$$\|b_t(\theta_\star)\| \le C h_t^p, \qquad \|b_t(\theta) - b_t(\theta_\star)\| \le C h_t^p \|\theta - \theta_\star\|, \qquad h_t \asymp t^{-\kappa}, \tag{61}$$

and suppose the joint curvature block, including that of every auxiliary joint path used in (D4), satisfies

$$\mathbb{E}\|\widehat{\mathrm{H}}_t^\star\|_{\mathrm{F}}^2 \le C \left\{ 1 + (R_t \delta_t^q)^{-1} \right\}, \quad R_t \asymp t^r, \quad \delta_t \asymp t^{-\kappa_{\mathrm{H}}}. \tag{62}$$

If

$$p\kappa > \frac{1}{2}, \qquad 2\gamma - (q\kappa_{\mathrm{H}} - r)_+ > 1 + \alpha, \tag{63}$$

where $x_+ = \max(x, 0)$, *then the bias conditions* (51) *and* (52), *as well as the joint stabilization condition, hold. Under the remaining conditions of Theorem 3.4, the joint bootstrap is valid. If, in addition, the frozen path satisfies the oracle, rate, localization, and original-path conditions of Corollary C.8, then the frozen bootstrap is also valid and paired frozen and joint replicas using the same primary multipliers are first-order equivalent.*

For the common choice $\lambda_t \asymp t^{-1}$, the curvature condition reduces to $(q\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha$. The first inequality in (63) is the usual root-n undersmoothing condition for a score bias of order h_t^p. At the boundary $p\kappa = 1/2$, the original average generally has a nonzero limiting mean shift that a centered multiplier bootstrap does not remove; Section C.8 gives the correction explicitly.

Corollary C.11. *Use the direct quantile score* (45) *as the primary oracle and the common-observation score secant* (24) *for the preconditioner. Let* $\eta_t = \eta_0 t^{-\alpha}$ *with* $1/2 < \alpha < 1$. *Suppose conditions (i) to (iii) of Corollary C.5 hold with* $\alpha(1+\nu) > 1$, *the original, frozen, and joint parameter paths satisfy (D4),* $\mathbb{E}\|\mathrm{M}_1\|_{\mathrm{F}}^2 < \infty$, *and the original preconditioner satisfies* (54). *Let*

$$\mathbb{E}\|\widehat{\mathrm{H}}_t^\star\|_{\mathrm{F}}^2 \le C\{1 + (R_t \delta_t)^{-1}\}, \qquad \lambda_t \asymp t^{-1}, \quad R_t \asymp t^r, \quad \delta_t \asymp t^{-\kappa_{\mathrm{H}}}. \tag{64}$$

All positive and negative score evaluations formed from one auxiliary observation share one multiplier. If

$$(\kappa_{\mathrm{H}} - r)_+ < 1 - \alpha, \tag{65}$$

then both the frozen and joint bootstraps are conditionally valid for

$$N\{0, \mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}\},$$

and paired frozen and joint replicas are first-order equivalent when they use the same primary multipliers.

The second-moment order in (64) is the characteristic indicator crossing rate: one score secant has variance of order δ_t^{-1}, and averaging R_t independent auxiliary blocks divides that component by R_t. For example, suppose the conditional density is bounded by $\bar{f}$ at every query location and U is independent of (X, Y), with $\mathbb{E}(\|X\|^3\|U\|^3) < \infty$. A single symmetrized secant then satisfies

$$\mathbb{E}\|\mathrm{H}_\delta\|_{\mathrm{F}}^2 \le \frac{\bar{f}}{2\delta}\mathbb{E}\{\|X\|^2\|U\|^2|X^\mathsf{T}U|\} \le \frac{\bar{f}}{2\delta}\mathbb{E}(\|X\|^3\|U\|^3).$$

Its mean is uniformly bounded, and averaging R_t independent observation blocks gives the stated bound; independent finite-variance observation weights only change the constant. Separate third moments suffice under independence of X and U. The normalization $\mathbb{E}(UU^\mathsf{T}) = \mathrm{I}_d$ alone does not supply this requirement. The density bound here is stronger than local density control at the true quantile and is only a sufficient example for the separate secant-moment assumption. Stating the moment bound directly avoids imposing globally bounded regressors or directions. The parameter-rate requirement remains separate and must be verified for the chosen secant recursion.

C.2.4 Invalid weighting of multiple evaluations

Proposition C.12. *Fix Z, U, θ, let $L_+ = L(\theta + hU; Z)$ and $L_- = L(\theta - hU; Z)$, and let W, W^+, W^- be independent copies satisfying* (12)*. Conditional on (Z, U), the perturbation covariance of the valid weighted difference is*

$$\mathrm{Cov}^\star\left((W-1)\frac{L_+ - L_-}{2h}U\right) = \frac{(L_+ - L_-)^2}{4h^2}UU^\mathsf{T}, \tag{66}$$

whereas independent side weights give

$$\mathrm{Cov}^\star\left(\frac{(W^+-1)L_+ - (W^- - 1)L_-}{2h}U\right) = \frac{L_+^2 + L_-^2}{4h^2}UU^\mathsf{T}. \tag{67}$$

If

$$\mathbb{E}\left[\{L_+ - L(\theta; Z)\}^2\|U\|^2 + \{L_- - L(\theta; Z)\}^2\|U\|^2\right] \longrightarrow 0$$

and $0 < \mathbb{E}\{L(\theta; Z)^2\|U\|^2\} < \infty$, the trace of the expectation of (67) *is of order h^{-2} as $h \to 0$, while* (66) *remains bounded under an L^2 local-Lipschitz condition.*

Corollary C.13. *Adding any random or deterministic term $C(Z)$ that does not depend on θ leaves the valid difference and* (66) *unchanged. Under the replacement $L_\pm \mapsto L_\pm + C(Z)$, the numerator of the invalid covariance changes by $2C(Z)\{L_+ + L_-\} + 2C(Z)^2$. Thus the invalid construction is not invariant to loss-level shifts in general. In particular, a sufficiently large nonzero constant shift changes its conditional covariance whenever $U \neq 0$. Failure of this invariance is a finite-sample diagnostic of invalid side weighting.*

Proposition C.14. *Let $D_1, \ldots, D_m \in \mathbb{R}^d$ be direction contributions computed from one observation. A common observation-level multiplier gives conditional perturbation covariance*

$$\mathrm{Cov}^\star\left\{(W-1)\sum_{j=1}^m D_j\right\} = \sum_{j=1}^m\sum_{k=1}^m D_j D_k^\mathsf{T}. \tag{68}$$

Independent direction multipliers give

$$\mathrm{Cov}^\star\left\{\sum_{j=1}^m (W_j - 1)D_j\right\} = \sum_{j=1}^m D_j D_j^\mathsf{T}.$$

Thus all cross terms $D_j D_k^\mathsf{T}$, $j \neq k$, are removed unless they vanish in the original covariance.

These propositions are exact finite-sample calculations, not asymptotic approximations. They apply to scalar finite differences, score secants, antithetic query pairs, and multi-direction gradient estimators. A faithful observation-level perturbation keeps all evaluations derived from one auxiliary observation in one weighted block. Separate weights generally change the target covariance, except when the omitted cross terms vanish or the affected component is first-order negligible; independent weights remain appropriate across genuinely independent auxiliary observations.

C.2.5 Inference with fixed or increasing numbers of replicas

The paired linear representations yield two distinct inferential regimes. A fixed number of replicas is enough for Student or normal-Wishart F pivots. Estimating the full conditional law or its covariance by an empirical distribution requires the number of replicas to increase.

The main statement is Theorem 3.5.

The theorem converts conditional bootstrap consistency into the asymptotic independence needed for fixed-replica pivots. The scalar calibration in the next corollary is the cheap-bootstrap construction (Lam & Wang, 2026); the theorem verifies its joint-limit premise for the additional recursions considered here.

Corollary C.15. *Under Theorem 3.5, let $a \in \mathbb{R}^d$ satisfy $a^{\mathsf{T}}\mathrm{V}a > 0$. For every fixed $B \geq 1$,*

$$\frac{a^{\mathsf{T}}(\bar{\theta}_n - \theta_\star)}{s_{n,B}(a)} \Longrightarrow t_B.$$

The ratio is evaluated on $\{s_{n,B}(a) > 0\}$, whose probability converges to one. Consequently, the interval in (26) *has limiting coverage $1 - \zeta$.*

Corollary C.16. *Under Theorem 3.5, let $\mathrm{C} \in \mathbb{R}^{q\times d}$ have full row rank and suppose $\mathrm{CVC}^{\mathsf{T}}$ is positive definite. For fixed $B \geq q$, define*

$$H_{n,B}(\mathrm{C}) = \begin{cases} (\mathrm{C}\bar{\theta}_n - \mathrm{C}\theta_\star)^{\mathsf{T}}\widehat{\Sigma}_{n,B}(\mathrm{C})^{-1}(\mathrm{C}\bar{\theta}_n - \mathrm{C}\theta_\star), & \mathcal{R}_{n,B}(\mathrm{C}), \\ +\infty, & \mathcal{R}_{n,B}(\mathrm{C})^c. \end{cases}$$

Then

$$H_{n,B}(\mathrm{C}) \Longrightarrow \frac{qB}{B-q+1}F_{q,B-q+1}, \qquad \mathbb{P}\{\mathcal{R}_{n,B}(\mathrm{C})^c\} \longrightarrow 0. \tag{69}$$

Hence $\mathcal{E}_{n,B}(\mathrm{C})$ has limiting coverage $1 - \zeta$.

Corollary C.17. *Under the conditions of Theorem 3.1, let $a \in \mathbb{R}^d$ have positive limiting variance and suppose the limiting scalar distribution is continuous. The ideal conditional interval* (28) *satisfies*

$$\mathbb{P}\{a^{\mathsf{T}}\theta_\star \text{ belongs to } (28)\} \longrightarrow 1 - \zeta.$$

If $B = B_n \to \infty$ and the implemented replicas are conditionally i.i.d., replacing the two ideal conditional quantiles by their empirical counterparts preserves the same limiting coverage.

Proposition C.18. *Under the conditions of Theorem 3.1, assume conditional uniform integrability of $\|T_n\|^2$, meaning that, for every $\varepsilon > 0$,*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \mathbb{P}\left[\mathbb{E}^\star\{\|T_n\|^2 \mathbf{1}(\|T_n\| > K)\} > \varepsilon\right] = 0. \tag{70}$$

Let $B = B_n \to \infty$ and let the implemented replicas be conditionally i.i.d. Then

$$\widehat{\mathrm{V}}_{n,B_n} \xrightarrow{p} \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}. \tag{71}$$

Corollary C.19. *Let $a_1, \ldots, a_q$ be fixed, with all limiting contrast variances positive. Under the conditions of Theorem 3.1 and Proposition C.18, let $\widehat{c}_{1-\zeta,n,B_n}$ be the empirical $(1-\zeta)$ quantile, across the B_n replicas, of* (31)*. If any $a_j^{\mathsf{T}}\widehat{\mathrm{V}}_{n,B_n}a_j$ is nonpositive, define the simultaneous output to be empty. The probability of this fallback event tends to zero, and otherwise the intervals* (32) *with $B = B_n$ have limiting familywise coverage $1 - \zeta$.*

The two regimes answer different computational questions. Fixed B gives asymptotically exact scalar intervals, but its Student critical value can be large and it cannot consistently estimate V.

Increasing B supports covariance estimation, conditional quantiles, and simultaneous maxima at a cost proportional to the number of paths. For a q-dimensional ellipsoid, $B \geq q$ is required; at $B = q$, the limiting F law has one denominator degree of freedom. The theorem uses exact rank. A scale-invariant numerical rank threshold preserves the limit if it tends to zero; a fixed tolerance is only a finite-precision safeguard. A numerical rank failure should return the empty region and count as noncoverage.

C.2.6 Relation to the cheap online bootstrap

When $\beta = 0$, $\mathrm{P} = \mathrm{I}_d$, and $W_t \sim \mathrm{Exp}(1)$, the original and weighted recursions in Section 2.3 coincide with the Cheap Online Bootstrap (COnB) (Lam & Wang, 2026), provided initialization, step-size indexing, and the averaging convention are matched. The scalar interval (26) is their t_B interval. The overlap is therefore exact at the algorithm and fixed-replica scalar-calibration levels for ordinary averaged SGD; empirical quantiles computed from many of the same paths are a separate increasing-replica calibration, not a different algorithm.

Lam and Wang establish fixed-B validity from primitive smooth strongly convex assumptions, analyze a cheap offline bootstrap, and derive refined Berry-Esseen bounds. They also give a separate two-stage procedure for sparse linear regression after support selection. The comparison here is restricted to $\alpha \in (1/2, 1)$; it does not use the $\alpha = 1$ endpoint in their printed assumptions. The fixed-momentum and dynamic-preconditioning theorems here use $\alpha \in (1/2, 1)$ and are first-order asymptotic results. They do not reproduce the offline method, the sparse-model-selection guarantee, or the finite-sample bounds (Lam & Wang, 2026).

The additional scope of the present analysis is algorithmic and local: fixed normalized momentum, locally nonsmooth estimating equations, predictable dynamic preconditioning, jointly updated secant states, and approximate scores. This scope requires path-stability and localization conditions for the chosen algorithm. For smooth ordinary averaged SGD with a small fixed number of replicas, Lam & Wang (2026) provide stronger direct guarantees; the results here address the additional recursions above. Neither theory contains the other.

C.3 Applications

This section verifies the derivative, innovation covariance, and regularity conditions in common stochastic optimization and statistical estimation problems. Whenever the stated algorithmic assumptions hold, each limiting law can be used with the fixed-replica intervals in Corollaries C.15 and C.16 or the empirical-quantile interval in Corollary C.17. Covariance and simultaneous output in Proposition C.18 and Corollary C.19 additionally require the conditional uniform-integrability condition (70); the application corollaries below do not assert it automatically. The examples are fixed-dimensional and concern a locally identified solution. Global convexity is one route to path stability, but the linearization itself is local once the conditions in Assumptions C.3 and C.6 have been verified.

In Table 14, $R(\theta) = \mathbb{E}\ell(\theta; Z)$ is the population risk, $\varepsilon = Y - X^{\mathsf{T}}\theta_\star$ is the least-squares residual, $\pi_\star = \pi(X^{\mathsf{T}}\theta_\star)$ with $\pi(u) = (1 + e^{-u})^{-1}$, and $s = \nabla_\theta \ell(\theta_\star; Z)$ is the target loss gradient.

Table 14: Identification of the derivative and innovation covariance in the applications.

Problem	Derivative A	Innovation covariance Q	Limiting covariance of $\sqrt{n}(\bar{\theta}_n - \theta_\star)$
Smooth M-estimation	$\mathrm{A} = \nabla^2 R(\theta_\star)$	$\mathrm{Q} = \mathbb{E}\{GG^{\mathsf{T}}\}$ at $\theta_\star$	$\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$; A^{-1} under the information identity
Least squares	$\mathrm{A} = \mathbb{E}(XX^{\mathsf{T}})$	$\mathrm{Q} = \mathbb{E}(\varepsilon^2 XX^{\mathsf{T}})$	$\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$; $\sigma^2\mathrm{A}^{-1}$ under conditional homoskedasticity
Logistic regression	$\mathrm{A} = \mathbb{E}\{\pi_\star(1 - \pi_\star)XX^{\mathsf{T}}\}$	A if correctly specified; otherwise Q in (80)	A^{-1} if correctly specified; otherwise $\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$

Problem	Derivative A	Innovation covariance Q	Limiting covariance
Quantile regression	$\mathrm{A}_\tau = \mathbb{E}\{f_{Y\mid X}(X^\mathsf{T}\theta_\tau \mid X)XX^\mathsf{T}\}$	S_τ in (49)	$\mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}$
Central random-direction difference	$\mathrm{A} = \nabla^2 R(\theta_\star)$	$\mathrm{Q}_U = \mathbb{E}\{UU^\mathsf{T}ss^\mathsf{T}UU^\mathsf{T}\}$	$\mathrm{A}^{-1}\mathrm{Q}_U\mathrm{A}^{-\mathsf{T}}$

C.3.1 Smooth stochastic M-estimation

Let $Z_1, Z_2, \ldots$ be i.i.d., let $R(\theta) = \mathbb{E}\ell(\theta; Z)$, and suppose $\theta_\star$ is a locally unique solution of $\nabla R(\theta) = 0$. Put

$$\mathrm{A} = \nabla^2 R(\theta_\star), \qquad \mathrm{Q} = \mathbb{E}\{G(\theta_\star, Z)G(\theta_\star, Z)^\mathsf{T}\}, \qquad G(\theta, Z) = \nabla_\theta \ell(\theta; Z).$$

Corollary C.20. *Suppose* A *is nonsingular and, for some* $\nu \in (0, 1]$ *and* $\chi \in (0, 2]$,

$$\nabla R(\theta_\star + e) = \mathrm{A}e + O(\|e\|^{1+\nu}), \tag{72}$$

$$\mathbb{E}\|\{G(\theta, Z) - \nabla R(\theta)\} - \{G(\theta_\star, Z) - \nabla R(\theta_\star)\}\|^2 \le C\|\theta - \theta_\star\|^\chi, \tag{73}$$

$$\mathbb{E}\|G(\theta_\star, Z)\|^2 < \infty. \tag{74}$$

Let $\eta_t = \eta_0 t^{-\alpha}$ *with* $1/2 < \alpha < 1$ *and* $\alpha(1 + \nu) > 1$. *If either*

(a) the fixed-momentum recursion satisfies the stability and localization conditions in Assumption C.3, or

(b) the dynamically preconditioned recursion has $b_t \equiv 0$ *and satisfies Assumptions C.6 and C.7,*

then

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}), \tag{75}$$

and the corresponding multiplier bootstrap consistently estimates this law.

A common primitive condition for (73) is a random local Lipschitz envelope:

$$\|G(\theta, Z) - G(\theta_\star, Z)\| \le L(Z)\|\theta - \theta_\star\|, \qquad \mathbb{E}L(Z)^2 < \infty.$$

Then $\chi = 2$. A locally Hölder Hessian of the population risk gives (72). These conditions involve only the first derivative of the sample loss and the local derivative of its expectation; a uniformly bounded sample Hessian is not required. For a correctly specified likelihood, the information identity gives $\mathrm{Q} = \mathrm{A}$ under the usual regularity conditions, and (75) reduces to $N(0, \mathrm{A}^{-1})$.

For joint dynamic preconditioning, consider the projected running Hessian

$$\mathrm{B}_t = \Pi_{\ell,u}\{\mathrm{B}_{t-1} + \lambda_t(\nabla^2 \ell(\theta_{t-1}; Z_t^H) - \mathrm{B}_{t-1})\}.$$

Here $Z_t^H = Z_{t1}^H$ denotes a curvature observation under the sampling and multiplier conventions in Section C.1.4, including reuse of the primary multiplier when $Z_t^H = Z_t$. If $\lambda_t \asymp t^{-1}$ and both the original and replica curvature observations have uniformly bounded second moments, then

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n-1} t^{\alpha/2}\lambda_t = O\{n^{(\alpha-1)/2}\} \longrightarrow 0.$$

Thus (57) holds for every $\alpha < 1$. The same calculation applies to a bounded-second-moment secant estimate. Parameter stability must still be verified for the chosen preconditioner.

C.3.2 LEAST SQUARES AND LOGISTIC REGRESSION

Corollary C.21. *Let (X_t, Y_t) be i.i.d. copies of (X, Y), and let $Y = X^\mathsf{T}\theta_\star + \varepsilon$, where $\mathbb{E}(X\varepsilon) = 0$, and define*

$$\mathrm{A} = \mathbb{E}(XX^\mathsf{T}), \qquad \mathrm{Q} = \mathbb{E}(\varepsilon^2 XX^\mathsf{T}).$$

Let $\eta_t = \eta_0 t^{-\alpha}$ with $1/2 < \alpha < 1$. Suppose A is positive definite and, for some $\delta \in (0, 2]$,

$$\mathbb{E}\|X\|^{2+\delta} < \infty, \qquad \mathbb{E}(\varepsilon^2\|X\|^2) < \infty, \qquad \left(1 + \frac{\delta}{2}\right)(1 + \alpha) > 2. \tag{76}$$

For fixed momentum, suppose the original and weighted parameter paths satisfy the global rate bounds in (M3), and the initial parameter and momentum states have finite $(1 + \delta/2)$ moments. For dynamic preconditioning without momentum, suppose the global original and weighted bounds in (53)*, finite initial parameter second moments, and Assumption C.7 hold, with $b_t \equiv 0$. Then*

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

and the online multiplier law is conditionally consistent. If $\mathbb{E}(\varepsilon^2 \mid X) = \sigma^2$, the covariance is $\sigma^2\mathrm{A}^{-1}$.

A fourth moment of X is not required by the following linearization argument, conditional on the stated parameter rates. Put $p = 1 + \delta/2 \in (1, 2]$. The proof controls the random linear coefficient $(XX^\mathsf{T} - \mathrm{A})(\theta_{t-1} - \theta_\star)$ in L^p by a martingale moment inequality rather than imposing quadratic-mean continuity in L^2. The relation $p(1 + \alpha) > 2$ is precisely the summability condition for that remainder. For α close to one, the design moment may therefore be only slightly above order two. This weaker moment condition does not by itself prove the assumed L^2 stability of the parameter path. For example, in scalar noiseless least squares with nonzero deterministic initial error and zero initial momentum, $e_1 = e_0\{1 - \eta_1(1 - \beta)\mathrm{P}X_1^2\}$ has infinite second moment when $\mathbb{E}X_1^4 = \infty$. Thus the assumed rate can fail even though the score moment used for linearization exists. Clipping the primary score changes the recursion and potentially its mean equation; such a modification requires separate analysis. When a raw sample Hessian is used in a joint curvature update, its second moment again corresponds to a fourth design moment; clipping the curvature observation separates that computational requirement from the primary-score inference. The mean field is exactly linear, so no Hölder remainder is needed. A sufficient local drift condition for a fixed or clipped dynamic preconditioner is $\mathrm{sym}(\mathrm{P}_t\mathrm{A}) \succeq \mu\mathrm{I}_d$ together with a uniform bound on $\|\mathrm{P}_t\|_{\mathrm{op}}$.

Corollary C.22. *Let (X_t, Y_t) be i.i.d. copies of (X, Y). Let $Y \mid X \sim \mathrm{Bernoulli}\{\pi(X^\mathsf{T}\theta_\star)\}$, where $\pi(v) = (1 + e^{-v})^{-1}$, and use $G(\theta; X, Y) = X\{\pi(X^\mathsf{T}\theta) - Y\}$. Suppose that, for some $\delta \in (0, 2]$,*

$$\mathbb{E}\|X\|^{2+\delta} < \infty, \qquad \nu_\delta = \min(\delta, 1), \qquad \alpha(1 + \nu_\delta) > 1, \tag{77}$$

and

$$\mathrm{A} = \mathbb{E}\left[\pi(X^\mathsf{T}\theta_\star)\{1 - \pi(X^\mathsf{T}\theta_\star)\}XX^\mathsf{T}\right] \tag{78}$$

is positive definite. Under the global original and weighted parameter-rate bounds and initial-moment conditions, and, for dynamic preconditioning, Assumption C.7,

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N(0, \mathrm{A}^{-1}), \tag{79}$$

and the multiplier bootstrap is conditionally valid. More generally, suppose $Y \in \{0, 1\}$, let $\theta_\star$ solve the possibly misspecified equation $\mathbb{E}[X\{\pi(X^\mathsf{T}\theta) - Y\}] = 0$, retain (77)*, and define*

$$\mathrm{Q} = \mathbb{E}\left[XX^\mathsf{T}\{Y - \pi(X^\mathsf{T}\theta_\star)\}^2\right]. \tag{80}$$

If A in (78) *is positive definite, then the limit is $N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1})$, and the same multiplier bootstrap is valid.*

Condition (77) can be substantially weaker than a fourth moment. It gives a population mean remainder of order $\|e\|^{1+\nu_\delta}$ and quadratic-mean score continuity of order $\|e\|^\delta$. Thus, for step sizes with α close to one, a moment only slightly above order two is sufficient. This moment condition is enough for fixed momentum and for any dynamic preconditioner whose stability and variation have been verified separately. A joint update based on an untrimmed sample Hessian may require an additional curvature moment; spectral truncation or a bounded score secant can avoid imposing that moment on the primary score.

C.3.3 NONSMOOTH REGRESSION

Quantile regression is covered by Corollaries C.5 and C.11. Under all conditions of the corresponding corollary, including identification and the original and weighted parameter rates,

$$\sqrt{n}(\bar{\theta}_n - \theta_\tau) \Longrightarrow N(0, \mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}), \tag{81}$$

where A_τ and S_τ are given by (48) and (49). Under the correctly specified linear conditional-quantile model $\Pr(Y \le X^\mathsf{T}\theta_\tau \mid X) = \tau$ almost surely, $\mathrm{S}_\tau = \tau(1-\tau)\mathbb{E}(XX^\mathsf{T})$, so the covariance in (81) becomes $\tau(1-\tau)\mathrm{A}_\tau^{-1}\mathbb{E}(XX^\mathsf{T})\mathrm{A}_\tau^{-1}$. The result uses the crossing probability of two indicator scores and does not require differentiability of the sample loss. The same proof pattern applies to other piecewise smooth scores when their changes are confined to events whose probability is of the same order as the parameter perturbation.

C.3.4 ZEROTH-ORDER STOCHASTIC OPTIMIZATION

Suppose only noisy loss evaluations are available. Let (Z_t, U_t) be i.i.d., with U_t independent of Z_t and $\mathbb{E}(U_tU_t^\mathsf{T}) = \mathrm{I}_d$, and form the common-observation central difference

$$\widehat{G}_{t,h_t}(\theta) = \frac{\ell(\theta + h_tU_t; Z_t) - \ell(\theta - h_tU_t; Z_t)}{2h_t}U_t. \tag{82}$$

This construction belongs to the Kiefer-Wolfowitz and simultaneous-perturbation family (Kiefer & Wolfowitz, 1952; Spall, 1992).

Corollary C.23. *Let $R(\theta) = \mathbb{E}\ell(\theta; Z)$, let $\mathrm{A} = \nabla^2 R(\theta_\star)$ be nonsingular, and write $s(Z) = \nabla_\theta \ell(\theta_\star; Z)$. Suppose the population gradient satisfies the local mean condition in* (72)*, and, uniformly for θ in the localization neighborhood as $h \downarrow 0$,*

$$\|\mathbb{E}\widehat{G}_h(\theta) - \nabla R(\theta)\| \le Ch^2, \tag{83}$$

$$\|\{\mathbb{E}\widehat{G}_h(\theta) - \nabla R(\theta)\} - \{\mathbb{E}\widehat{G}_h(\theta_\star) - \nabla R(\theta_\star)\}\| \le Ch^2\|\theta - \theta_\star\|, \tag{84}$$

$$\mathbb{E}\|\widehat{G}_h(\theta_\star) - UU^\mathsf{T}s(Z)\|^2 \le Ch^4, \tag{85}$$

and the centered finite-difference oracle has the quadratic-mean continuity required in (50)*. Assume*

$$\mathrm{Q}_U := \mathbb{E}\{UU^\mathsf{T}s(Z)s(Z)^\mathsf{T}UU^\mathsf{T}\} < \infty.$$

Let $h_t \asymp t^{-\kappa}$ with $\kappa > 1/4$, and suppose the no-momentum recursion with a fixed or dynamic preconditioner satisfies Assumptions C.6 and C.7. Both loss evaluations in (82) *receive the same multiplier. Then*

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N(0, \mathrm{A}^{-1}\mathrm{Q}_U\mathrm{A}^{-\mathsf{T}}), \tag{86}$$

and the multiplier bootstrap is conditionally valid. If $U \sim N(0, \mathrm{I}_d)$, then

$$\mathrm{Q}_U = 2\mathrm{S} + \mathrm{tr}(\mathrm{S})\mathrm{I}_d, \qquad \mathrm{S} = \mathbb{E}\{s(Z)s(Z)^\mathsf{T}\}. \tag{87}$$

The bias and L^2 approximation conditions in (83) and (85) follow from a third-order directional Taylor expansion with a square-integrable remainder envelope valid along all queried segments, including those generated by unbounded directions. The local bias-increment condition (84) is additional; it follows, for example, from local $C^{3,1}$ smoothness with an integrable derivative envelope. The conditions are stated directly to avoid imposing a particular high moment on the direction distribution. The threshold $\kappa > 1/4$ is the root-n undersmoothing condition for a central-difference bias of order h_t^2. The limit in (86) differs from the gradient-based sandwich because random directions add the covariance Q_U. A faithful observation-level perturbation uses one multiplier for the two loss evaluations; separate side weights generally change the target covariance and, under Proposition C.12, make it diverge.

C.4 CONDITIONAL TRANSFER TOOLS

Lemma C.24. *If $R_n = o_{\mathbb{P}\otimes\mathbb{P}^\star}(1)$, then for every $\varepsilon > 0$,*

$$\mathbb{P}^\star(\|R_n\| > \varepsilon) \xrightarrow{p} 0. \tag{88}$$

Equivalently, $R_n = o_{\mathbb{P}^\star}(1)$ in $\mathbb{P}$-probability.

Lemma C.25. *Under Assumption C.1,*

$$d_{\mathrm{BL}}\left[\mathcal{L}^\star\left\{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(W_t-1)\psi_{nt}\right\}, N(0,\mathrm{Q})\right] \xrightarrow{p} 0.$$

Lemma C.26. *Let X_n be measurable with respect to the original data and suppose $X_n \Longrightarrow X$. Conditional on the original data, let $Y_{n1},\dots,Y_{nB}$ be independent, where B is fixed. If, for every $b=1,\dots,B$,*

$$d_{\mathrm{BL}}\{\mathcal{L}^\star(Y_{nb}),\mathcal{L}(Y)\} \xrightarrow{p} 0$$

for a nonrandom probability law $\mathcal{L}(Y)$, then

$$(X_n, Y_{n1},\dots,Y_{nB}) \Longrightarrow (X, Y_1,\dots,Y_B),$$

where $Y_1,\dots,Y_B$ are independent copies of Y and are independent of X.

Lemma C.27. *Let $\psi_1,\psi_2,\dots$ be i.i.d. with $\mathbb{E}\psi_t = 0$ and $\mathbb{E}\|\psi_t\|^2 < \infty$. With $\psi_{nt} = \psi_t$, Assumption C.1 holds and $\mathrm{Q} = \mathbb{E}(\psi_t\psi_t^{\mathsf{T}})$.*

C.5 Fixed-momentum tools

Lemma C.28. *Let $\beta \in [0,1)$ and $m_t = \beta m_{t-1} + (1-\beta)Y_t$. If $\mathbb{E}\|m_0\|^2 < \infty$ and $\sup_t \mathbb{E}\|Y_t\|^2 < \infty$, then $\sup_t \mathbb{E}\|m_t\|^2 < \infty$.*

Lemma C.29. *Let $\eta_t = \eta_0 t^{-\alpha}$ with $0 < \alpha < 1$, let $\mathbb{E}\|u_0\| < \infty$, and suppose $\mathbb{E}\|u_t\| \le C\eta_t^{1/2}$ for $t \ge 1$. Then*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{u_{t-1}-u_t}{\eta_t} = o_{L^1}(1).$$

C.6 Dynamic-preconditioning tools

We first record the summation-by-parts estimate that isolates the dynamic preconditioner. It is stated under a generic probability law and therefore applies to both the original and weighted paths.

Lemma C.30. *Let $a_t = \eta_t^{-1} = \eta_0^{-1}t^\alpha$, with $0 < \alpha < 1$. Suppose $\mathbb{E}\|e_0\|^2 < \infty$, $\mathbb{E}\|e_t\|^2 \le C\eta_t$, $\mathbb{E}\|\mathrm{M}_1\|_{\mathrm{F}}^2 < \infty$, and*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n-1} t^{\alpha/2}\{\mathbb{E}\|\mathrm{M}_{t+1}-\mathrm{M}_t\|_{\mathrm{F}}^2\}^{1/2} \longrightarrow 0.$$

Then

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} a_t \mathrm{M}_t(e_{t-1}-e_t) = o_p(1).$$

The same conclusion holds under the joint law when the assumptions hold under that law.

C.6.1 One primitive route to the parameter-rate condition

The $O(\eta_t)$ rate in (53) is deliberately separated from preconditioner stabilization because it is an optimization property. The following standard drift condition gives one directly checkable sufficient route. It is not necessary and need not hold for all secant constructions. For a continuous biased mean and an invertible preconditioner, the strict drift below forces $b_t(\theta_\star) = 0$: substitution of $e = \pm sv$ and passage to $s \downarrow 0$ eliminates any nonzero linear term. Thus this primitive directly covers exact scores, rather than every approximate-score construction.

Proposition C.31. *Consider either the original dynamic recursion or a weighted recursion with predictable P_t. Suppose globally that, for $e = \theta - \theta_\star$,*

$$e^{\mathsf{T}}\mathrm{P}_t\{g(\theta)+b_t(\theta)\} \ge \mu\|e\|^2, \tag{89}$$

$$\|\mathrm{P}_t\|_{\mathrm{op}} \le \bar{p},$$

$$\mathbb{E}\{\|\widehat{G}_t(\theta)\|^2 \mid \mathcal{F}_{t-1}\} \le C(1+\|e\|^2) \tag{90}$$

for fixed $\mu, \bar{p} > 0$. Assume $\mathbb{E}\|\theta_0-\theta_\star\|^2 < \infty$ and let $\eta_t = \eta_0 t^{-\alpha}$ with $0 < \alpha < 1$. Then, after a finite initial segment, $\mathbb{E}\|\theta_t-\theta_\star\|^2 \le C\eta_t$. The same conclusion holds for a weighted path with a finite initial second moment when the conditional mean of the weighted oracle is $g + b_t$ and its conditional second moment satisfies (90) *with another constant.*

For an explicit global example, take $g(\theta_\star + e) = e$ with additive centered finite-variance noise and a predictable symmetric preconditioner satisfying $p_- \mathrm{I}_d \preceq \mathrm{P}_t \preceq p_+ \mathrm{I}_d$ for fixed $0 < p_- \le p_+ < \infty$. The drift and moment bounds hold for both paths because the current mean-one multiplier is independent of the pre-update history. If the preconditioner comes from a projected curvature update with gain $O(t^{-1})$ and uniformly bounded curvature second moments, Proposition C.9 also verifies its variation condition. This closes both requirements in a concrete class without requiring curvature consistency.

C.7 Application tools

Lemma C.32. *Let $1 < p \le 2$ and let $D_1, \dots, D_n$ be an $\mathbb{R}^d$-valued martingale-difference sequence with $\mathbb{E}\|D_t\|^p < \infty$. Then*

$$\mathbb{E}\left\|\sum_{t=1}^{n} D_t\right\|^p \le C_{p,d} \sum_{t=1}^{n} \mathbb{E}\|D_t\|^p. \tag{91}$$

C.8 Finite-difference bias at the root-n boundary

The centered multiplier bootstrap reproduces stochastic fluctuations, but a deterministic score bias at the root-n boundary must be removed from the center of the original estimator. The next calculation makes the issue explicit.

Proposition C.33. *Consider the original and weighted no-momentum recursions of Theorem 3.3, allowing a fixed preconditioner. Retain all its assumptions except the first limit in* (51)*. Instead, suppose $p > 0$, $h_0 > 0$, b_{lead} is a fixed vector, and $h_t = h_0 t^{-1/(2p)}$, with*

$$b_t(\theta_\star) = b_{\mathrm{lead}} h_t^p + r_{b,t}, \qquad \frac{1}{\sqrt{n}} \sum_{t=1}^{n} r_{b,t} \xrightarrow{p} 0, \qquad \frac{1}{n} \sum_{t=1}^{n} \|r_{b,t}\|^2 \xrightarrow{p} 0. \tag{92}$$

Then

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N\left(-2h_0^p \mathrm{A}^{-1} b_{\mathrm{lead}},\ \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}\right). \tag{93}$$

The centered multiplier statistic $\sqrt{n}(\bar{\theta}_n^\star - \bar{\theta}_n)$ still has the zero-mean Gaussian limit and therefore misses the shift in (93)*.*

If estimates $\widehat{\mathrm{A}}$ and $\widehat{b}_t$ satisfy

$$\widehat{\mathrm{A}} \xrightarrow{p} \mathrm{A}, \qquad \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{\widehat{b}_t - b_t(\theta_\star)\} = o_p(1), \tag{94}$$

then

$$\bar{\theta}_n^{\mathrm{bc}} = \bar{\theta}_n + \widehat{\mathrm{A}}^{-1} \frac{1}{n} \sum_{t=1}^{n} \widehat{b}_t \tag{95}$$

has the centered Gaussian limit, and the same centered multiplier law may be used around $\bar{\theta}_n^{\mathrm{bc}}$. The bootstrap fluctuations remain $\sqrt{n}(\bar{\theta}_n^\star - \bar{\theta}_n)$, and their quantiles center intervals at $\bar{\theta}_n^{\mathrm{bc}}$; uncorrected replica averages are not recentered at $\bar{\theta}_n^{\mathrm{bc}}$.

Alternatively, if an oracle at scale h has conditional mean expansion

$$\mathbb{E}\{\widehat{G}_{t,h}(\theta) \mid \mathcal{F}_{t-1}\} = g(\theta) + c_t(\theta) h^p + O(h^{p+\delta_b}) \tag{96}$$

uniformly locally for some $\delta_b > 0$, the Richardson block

$$\widetilde{G}_{t,h}(\theta) = \frac{2^p \widehat{G}_{t,h/2}(\theta) - \widehat{G}_{t,h}(\theta)}{2^p - 1} \tag{97}$$

removes the h^p term. All evaluations in (97) *formed from the same observation must share one observation-level multiplier. This calculation establishes cancellation of the leading bias term. Centered inference for the Richardson recursion additionally requires the remaining $O(h_t^{p+\delta_b})$ bias and its local increments to satisfy the bias, quadratic-mean continuity, and path-stability conditions of the underlying linearization theorem.*

D PROOFS

Proof of Lemma C.24. Fix $\varepsilon > 0$ and $\eta > 0$. The random variable $\mathbb{P}^\star(\|R_n\| > \varepsilon)$ takes values in $[0, 1]$. Markov's inequality and iterated expectation give

$$\begin{aligned}\mathbb{P}\left\{\mathbb{P}^\star(\|R_n\| > \varepsilon) > \eta\right\} &\le \eta^{-1}\mathbb{E}\left[\mathbb{P}^\star(\|R_n\| > \varepsilon)\right] \\ &= \eta^{-1}(\mathbb{P} \otimes \mathbb{P}^\star)(\|R_n\| > \varepsilon).\end{aligned}$$

The final expression converges to zero by the joint-probability assumption. Since η is arbitrary, (88) follows. □

Proof of Lemma C.25. We verify the conditional Lindeberg-Feller theorem by the Cramér-Wold device.

Step 1: conditional centering and variance. Fix $a \in \mathbb{R}^d$ and define $X_{nt} = n^{-1/2}(W_t - 1)a^\mathsf{T}\psi_{nt}$. Conditional on the original data, these variables are independent and centered. Their conditional variance sum is

$$\sum_{t=1}^{n} \mathbb{E}^\star\{(X_{nt})^2\} = a^\mathsf{T}\left(\frac{1}{n}\sum_{t=1}^{n}\psi_{nt}\psi_{nt}^\mathsf{T}\right)a \xrightarrow{p} a^\mathsf{T}\mathrm{Q}a. \tag{98}$$

Step 2: conditional Lindeberg condition. For $a \neq 0$ and every $\varepsilon > 0$,

$$\begin{aligned}&\sum_{t=1}^{n}\mathbb{E}^\star\left[(X_{nt})^2\mathbf{1}\{|X_{nt}| > \varepsilon\}\right] \\ &\quad\le \frac{\|a\|^2}{n}\sum_{t=1}^{n}\|\psi_{nt}\|^2\mathbb{E}\left[(W_t-1)^2\mathbf{1}\left\{|W_t-1|\,\|\psi_{nt}\| > \frac{\varepsilon\sqrt{n}}{\|a\|}\right\}\right] \xrightarrow{p} 0\end{aligned} \tag{99}$$

by (34). The case $a = 0$ is immediate.

Step 3: conditional weak convergence. Take an arbitrary subsequence. By (98) and (99), it has a further subsequence along which both quantities converge almost surely for every a in a fixed countable dense subset of $\mathbb{R}^d$. The ordinary Lindeberg-Feller theorem, applied conditionally, gives

$$\sum_{t=1}^{n} X_{nt} \Longrightarrow N(0, a^\mathsf{T}\mathrm{Q}a) \quad \text{under } \mathbb{P}^\star$$

almost surely along that further subsequence. When $a^\mathsf{T}\mathrm{Q}a = 0$, the same conclusion follows from conditional Chebyshev's inequality. Conditional tightness follows from (98). The Cramér-Wold device therefore yields conditional weak convergence to $N(0, \mathrm{Q})$, and the bounded-Lipschitz distance converges to zero almost surely along the further subsequence. Since every subsequence admits such a further subsequence, the convergence holds in probability. □

Proof of Theorem 3.1. Subtract (36) from (37):

$$\sqrt{n}(\bar{\theta}_n^\star - \bar{\theta}_n) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(W_t - 1)\psi_{nt} + (r_n^\star - r_n). \tag{100}$$

By Lemma C.24, $r_n^\star = o_{\mathbb{P}^\star}(1)$ in $\mathbb{P}$-probability. Because r_n is measurable with respect to the original data and $r_n = o_p(1)$,

$$\mathbb{P}^\star(\|r_n\| > \varepsilon) = \mathbf{1}\{\|r_n\| > \varepsilon\} \xrightarrow{p} 0.$$

Hence $r_n^\star - r_n = o_{\mathbb{P}^\star}(1)$ in $\mathbb{P}$-probability.

By Lemma C.25 and continuity of the linear map $x \mapsto -\mathrm{A}^{-1}x$,

$$-\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(W_t - 1)\psi_{nt} \Longrightarrow N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}})$$

conditionally in probability. Conditional Slutsky applied to (100) proves (2).

For the original estimator, (36), (35), and ordinary Slutsky give

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) \Longrightarrow N(0, \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}). \tag{101}$$

The triangle inequality for d_{BL}, together with (2) and (101), proves (3). □

Proof of Lemma C.26. Let u and $v_1, \ldots, v_B$ be conformable vectors and write $\phi_{nb}(v) = \mathbb{E}^\star \exp(iv^\mathsf{T} Y_{nb})$. Bounded-Lipschitz convergence implies $\phi_{nb}(v) \xrightarrow{p} \phi_Y(v)$ for every fixed b and v. Since B is fixed and characteristic functions are bounded by one, each convergence also holds in L^1. Conditional independence gives

$$\begin{aligned}&\mathbb{E}\exp\left(iu^\mathsf{T} X_n + i\sum_{b=1}^{B} v_b^\mathsf{T} Y_{nb}\right)\\ &\quad= \mathbb{E}\left[\exp(iu^\mathsf{T} X_n)\prod_{b=1}^{B}\phi_{nb}(v_b)\right] \longrightarrow \phi_X(u)\prod_{b=1}^{B}\phi_Y(v_b).\end{aligned}$$

The limit is the characteristic function of independent $(X, Y_1, \ldots, Y_B)$. Lévy's continuity theorem proves the claim. □

Proof of Lemma C.27. The strong law of large numbers, applied entrywise to $\psi_t\psi_t^\mathsf{T}$, gives (33), and the multivariate i.i.d. central limit theorem gives (35). It remains to verify (34). By independence and identical distribution,

$$\mathbb{E}\{\mathcal{L}_n(\varepsilon)\} = \mathbb{E}\left[(W_1 - 1)^2\|\psi_1\|^2 \mathbf{1}\{|W_1 - 1|\,\|\psi_1\| > \varepsilon\sqrt{n}\}\right].$$

The integrand converges almost surely to zero and is dominated by $(W_1 - 1)^2\|\psi_1\|^2$, whose expectation is finite because the multiplier and influence are independent and both have finite second moments. Dominated convergence gives $\mathbb{E}\{\mathcal{L}_n(\varepsilon)\} \to 0$, hence $\mathcal{L}_n(\varepsilon) \xrightarrow{p} 0$ by Markov's inequality. □

Proof of Lemma C.28. Iteration gives

$$m_t = \beta^t m_0 + (1-\beta)\sum_{s=1}^{t}\beta^{t-s}Y_s.$$

The weights in the second term are nonnegative and sum to at most one. Jensen's inequality therefore gives

$$\mathbb{E}\|m_t\|^2 \le 2\beta^{2t}\mathbb{E}\|m_0\|^2 + 2(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\mathbb{E}\|Y_s\|^2,$$

which is uniformly bounded. □

Proof of Proposition C.4. We use one argument for either path. Put $\zeta_t = 1$ on the original path and $\zeta_t = W_t$ on the weighted path, and define

$$\widetilde{G}_t(e) = \zeta_t G(\theta_\star + e, Z_t), \qquad \varepsilon_t(e) = \widetilde{G}_t(e) - g(\theta_\star + e).$$

For a pre-update-predictable e, (41) and multiplier independence give

$$\mathbb{E}\{\varepsilon_t(e) \mid \mathcal{H}_{t-1}\} = 0, \qquad \mathbb{E}\{\|\varepsilon_t(e)\|^2 \mid \mathcal{H}_{t-1}\} \le C(1 + \|e\|^2). \tag{102}$$

The second bound follows from (42), global Lipschitz continuity of g, and $\mathbb{E}W_t^2 < \infty$.

Write $e_t = \theta_t - \theta_\star$, $c_\beta = \beta/(1-\beta)$, and $u_t = e_t - c_\beta\eta_t\mathrm{P}m_t$. For u_0, use η_1 in this definition. Direct substitution into the two momentum updates gives

$$u_t = u_{t-1} - \eta_t\mathrm{P}\widetilde{G}_t(e_{t-1}) + c_\beta(\eta_{t-1} - \eta_t)\mathrm{P}m_{t-1}.$$

Add and subtract $g(\theta_\star + u_{t-1})$ to obtain

$$u_t = u_{t-1} - \eta_t \mathrm{P} g(\theta_\star + u_{t-1}) - \eta_t \mathrm{P}\varepsilon_t(e_{t-1}) + d_t, \tag{103}$$

where

$$\begin{aligned} d_t = {} & c_\beta(\eta_{t-1} - \eta_t)\mathrm{P}m_{t-1} \\ & - \eta_t \mathrm{P}\{g(\theta_\star + e_{t-1}) - g(\theta_\star + u_{t-1})\}. \end{aligned}$$

Since $e_{t-1} - u_{t-1} = c_\beta \eta_{t-1}\mathrm{P}m_{t-1}$, the Lipschitz condition and $\eta_{t-1} - \eta_t = O(\eta_t/t)$ give, for all sufficiently large t,

$$\|d_t\| \le C\eta_t\eta_{t-1}\|m_{t-1}\|. \tag{104}$$

Here $t^{-1} = O(\eta_{t-1})$ because $\alpha < 1$.

All Euclidean and R-norms are equivalent. The drift and Lipschitz conditions imply

$$\|u - \eta_t \mathrm{P} g(\theta_\star + u)\|_{\mathrm{R}}^2 \le (1 - c\eta_t)\|u\|_{\mathrm{R}}^2$$

for some $c > 0$ and all sufficiently large t. In the conditional square of (103), the cross term involving $\varepsilon_t(e_{t-1})$ vanishes by (102). Young's inequality, (104), and

$$\|e_{t-1}\|^2 \le C\{\|u_{t-1}\|_{\mathrm{R}}^2 + \eta_{t-1}^2\|m_{t-1}\|^2\}$$

then yield

$$\mathfrak{u}_t \le (1 - c_1\eta_t)\mathfrak{u}_{t-1} + C\eta_t^2 + C\eta_t\eta_{t-1}^2\mathfrak{m}_{t-1}, \tag{105}$$

where $\mathfrak{u}_t = \mathbb{E}\|u_t\|_{\mathrm{R}}^2$ and $\mathfrak{m}_t = \mathbb{E}\|m_t\|^2$. For example, the perturbation contributes $C\eta_t^{-1}\mathbb{E}\|d_t\|^2$, while the conditional noise contributes at most $C\eta_t^2\{1+\mathfrak{u}_{t-1}+\eta_{t-1}^2\mathfrak{m}_{t-1}\}$; the $\mathfrak{u}_{t-1}$ term is absorbed into the contraction for large t.

Convexity of the squared norm in the momentum update gives

$$\begin{aligned} \mathfrak{m}_t &\le \beta\mathfrak{m}_{t-1} + (1-\beta)\mathbb{E}\|\widetilde{G}_t(e_{t-1})\|^2 \\ &\le \{\beta + C\eta_{t-1}^2\}\mathfrak{m}_{t-1} + C(1 + \mathfrak{u}_{t-1}). \end{aligned}$$

Choose $\bar\beta \in (\beta, 1)$. After increasing the deterministic starting index,

$$\mathfrak{m}_t \le \bar\beta\mathfrak{m}_{t-1} + C(1 + \mathfrak{u}_{t-1}). \tag{106}$$

For a fixed $K > 0$, set $\mathfrak{s}_t = \mathfrak{u}_t + K\eta_t^2\mathfrak{m}_t$. Combining (105) and (106), and using $\eta_t/\eta_{t-1} \to 1$, gives, after one further increase of the starting index,

$$\mathfrak{s}_t \le (1 - c_2\eta_t)\mathfrak{s}_{t-1} + C\eta_t^2 \tag{107}$$

for some $c_2 > 0$. Indeed, $KC\eta_t^2\mathfrak{u}_{t-1}$ is absorbed by the order-η_t contraction of $\mathfrak{u}_{t-1}$, while the coefficient of $K\eta_{t-1}^2\mathfrak{m}_{t-1}$ is bounded by

$$\bar\beta(\eta_t/\eta_{t-1})^2 + C\eta_t/K < 1 - c_2\eta_t$$

for large t.

The scalar recursion (107) implies $\mathfrak{s}_t = O(\eta_t)$. This follows by induction from $\eta_{t-1}/\eta_t = 1 + O(t^{-1})$ and $t^{-1} = o(\eta_t)$. Since $e_t = u_t + c_\beta\eta_t\mathrm{P}m_t$, norm equivalence gives

$$\mathbb{E}\|e_t\|^2 \le C\{\mathfrak{u}_t + \eta_t^2\mathfrak{m}_t\} = O(\eta_t).$$

The oracle second moment is now uniformly bounded along the path, so Lemma C.28 gives $\sup_t \mathfrak{m}_t < \infty$. Applying the same argument to $\zeta_t = 1$ and $\zeta_t = W_t$ proves (43) to (44). If the remaining mean and score conditions are global, no auxiliary localization is needed and $k_n = 0$. □

Proof of Lemma C.29. Set $a_t = \eta_t^{-1} = \eta_0^{-1}t^\alpha$. Summation by parts gives

$$\sum_{t=1}^{n} a_t(u_{t-1} - u_t) = a_1u_0 - a_nu_n + \sum_{t=1}^{n-1}(a_{t+1} - a_t)u_t. \tag{108}$$

By assumption, $\mathbb{E}\|u_t\| \le Ct^{-\alpha/2}$, $\mathbb{E}\|a_n u_n\| \le Cn^{\alpha/2}$. By the mean-value theorem, $a_{t+1} - a_t = O(t^{\alpha-1})$, and therefore

$$\mathbb{E}\left\|\sum_{t=1}^{n-1}(a_{t+1}-a_t)u_t\right\| \le C\sum_{t=1}^{n-1} t^{\alpha-1}t^{-\alpha/2} \le C\sum_{t=1}^{n-1} t^{\alpha/2-1} = O(n^{\alpha/2}).$$

The initial term is $O(1)$. Dividing (108) by $\sqrt{n}$ yields an L^1 bound of order $n^{(\alpha-1)/2} + n^{-1/2}$, which converges to zero because $\alpha < 1$. □

Proof of Theorem 3.2. We first work under global versions of the assumptions. Put $e_t = \theta_t - \theta_\star$ and $c_\beta = \beta/(1-\beta)$.

Step 1: an exact momentum-sum identity. For either path, write Y_t for the oracle entering its momentum update. Thus $Y_t = G(\theta_{t-1}, Z_t)$ on the original path and $Y_t = W_t G(\theta^\star_{t-1}, Z_t)$ on the weighted path. The momentum and parameter updates imply

$$Y_t = m_t + c_\beta(m_t - m_{t-1}), \qquad m_t = \mathrm{P}^{-1}\frac{e_{t-1}-e_t}{\eta_t}.$$

Consequently, on either path,

$$\sum_{t=1}^{n} Y_t = \mathrm{P}^{-1}\sum_{t=1}^{n}\frac{e_{t-1}-e_t}{\eta_t} + c_\beta(m_n - m_0). \tag{109}$$

Step 2: original representation. Write $\Delta\xi_t = \xi_t(\theta_{t-1}) - \xi_t(\theta_\star)$. Since

$$G(\theta_{t-1}, Z_t) = \mathrm{A}e_{t-1} + \psi_t + \Delta\xi_t + r(e_{t-1}),$$

(109) gives

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} e_{t-1} = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t + \mathrm{A}^{-1}\mathrm{P}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{e_{t-1}-e_t}{\eta_t} + c_\beta\mathrm{A}^{-1}\frac{m_n - m_0}{\sqrt{n}} - \mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{\Delta\xi_t + r(e_{t-1})\}.$$

The Abel term is $o_{L^1}(1)$ by Lemma C.29 and (39). The parameter update gives

$$\frac{1}{\sqrt{n}}\mathbb{E}\|m_n\| \le \frac{C}{\sqrt{n}\eta_n}\{\mathbb{E}\|e_{n-1}\|^2 + \mathbb{E}\|e_n\|^2\}^{1/2} = O\{n^{(\alpha-1)/2}\} = o(1),$$

and $m_0/\sqrt{n} = o_{L^1}(1)$. Martingale orthogonality, quadratic-mean continuity, and the iterate rate give

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\Delta\xi_t\right\|^2 \le \frac{C}{n}\sum_{t=1}^{n}\eta_{t-1}^{\chi/2} \longrightarrow 0.$$

Finally, because $1 + \nu \le 2$,

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} r(e_{t-1})\right\| \le \frac{C}{\sqrt{n}}\sum_{t=1}^{n}\eta_{t-1}^{(1+\nu)/2} \longrightarrow 0$$

by $\alpha(1+\nu) > 1$. Replacing the lagged average costs $(e_0 - e_n)/\sqrt{n} = o_p(1)$ and proves (4).

Step 3: weighted representation. Set $e_t^\star = \theta_t^\star - \theta_\star$ and $\Delta\xi_t^\star = \xi_t(\theta_{t-1}^\star) - \xi_t(\theta_\star)$. The exact expansion

$$W_t G(\theta^\star_{t-1}, Z_t) = \mathrm{A}e^\star_{t-1} + W_t\psi_t + W_t\Delta\xi_t^\star + (W_t - 1)\mathrm{A}e^\star_{t-1} + W_t r(e^\star_{t-1})$$

combined with (109) gives the same decomposition with leading term $-\mathrm{A}^{-1}n^{-1/2}\sum_t W_t\psi_t$. The weighted parameter rate and Lemma C.29 control the Abel term. The parameter update controls $(m_n^\star - m_0^\star)/\sqrt{n}$ exactly as on the original path. Under the enlarged filtration, the two arrays $W_t\Delta\xi_t^\star$ and $(W_t-1)\mathrm{A}e_{t-1}^\star$ are martingale differences. Their combined normalized sum has second moment bounded by

$$\frac{C}{n}\sum_{t=1}^{n}\{\eta_{t-1}^{\chi/2}+\eta_{t-1}\}\longrightarrow 0.$$

Also $\mathbb{E}|W_t| \le (\mathbb{E}W_t^2)^{1/2}$, so the nonlinear term is $o_{L^1}(1)$ by the same $\alpha(1+\nu)>1$ calculation. Replacing the lagged weighted average costs $(e_0^\star - e_n^\star)/\sqrt{n} = o_{\mathbb{P}\otimes\mathbb{P}^\star}(1)$, proving (5).

Step 4: localization and transfer. For local assumptions, apply Steps 1 to 3 to the auxiliary recursions in (M4). On their pathwise agreement events, the auxiliary and original root-n averages differ only through the first $k_n = o(\sqrt{n})$ states; the stated uniform second-moment bound makes this difference negligible. Because the global extensions agree with the original score on $\mathcal{N}$, the target influence sequence is unchanged. The two representations thus transfer to the original paths. The influence-limit premise is part of (M2), so Theorem 3.1 proves (6). □

Proof of Corollary C.5. Let $q_X = X^\mathsf{T}\theta_\tau$ and $s = X^\mathsf{T}e$. Conditional on X,

$$\mathbb{E}\{G_\tau(\theta_\tau+e;X,Y) - G_\tau(\theta_\tau;X,Y)\mid X\} = X\{F_{Y|X}(q_X+s\mid X) - F_{Y|X}(q_X\mid X)\}.$$

Step 1: local mean expansion. On $|s|\le r_0$, integration of (46) gives

$$F_{Y|X}(q_X+s\mid X) - F_{Y|X}(q_X\mid X) = f_{Y|X}(q_X\mid X)s + R_X(s), \qquad |R_X(s)|\le CL(X)|s|^{1+\nu}.$$

On $|s|>r_0$, the absolute remainder is bounded by $1+F(X)|s|$. Since $|s|\le\|X\|\|e\|$, the event $|s|>r_0$ is contained in $\{\|X\|>r_0/\|e\|\}$. The local part is bounded by

$$C\|e\|^{1+\nu}\mathbb{E}\{L(X)\|X\|^{2+\nu}\}.$$

For the two tail terms, Markov's inequality yields

$$\mathbb{E}\left[\|X\|\mathbf{1}\{\|X\|>r_0/\|e\|\}\right] \le C\|e\|^{1+\nu}\mathbb{E}\|X\|^{2+\nu},$$
$$\|e\|\mathbb{E}\left[F(X)\|X\|^2\mathbf{1}\{\|X\|>r_0/\|e\|\}\right] \le C\|e\|^{1+\nu}\mathbb{E}\{F(X)\|X\|^{2+\nu}\}.$$

All expectations are finite by (47). The bounds hold for every $e\in\mathbb{R}^d$, since the split imposes no upper bound on $\|e\|$. Therefore

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

Step 2: quadratic-mean continuity. The two score indicators differ only when Y lies between q_X and q_X+s. Suppressing the arguments (X,Y), we therefore have

$$\begin{aligned}\mathbb{E}\|G_\tau(\theta_\tau+e)-G_\tau(\theta_\tau)\|^2 &\le \mathbb{E}\left[\|X\|^2\{1\wedge(F(X)|s|)\}\mathbf{1}\{|s|\le r_0\}\right] + \mathbb{E}\left[\|X\|^2\mathbf{1}\{|s|>r_0\}\right]\\ &\le C\|e\|^\nu\mathbb{E}\{(1+F(X)^\nu)\|X\|^{2+\nu}\}\le C\|e\|^\nu.\end{aligned}$$

Here we used $1\wedge x\le x^\nu$ for $x\ge 0$ and $F(X)^\nu\le 1+F(X)$. Centering changes the bound by at most a constant factor, so the score has quadratic-mean exponent $\chi=\nu$.

Step 3: remaining conditions. The target score is bounded in norm by $\|X\|$, and (47) implies its finite second moment. Thus Lemma C.27 applies without a moment above order two. Positive definiteness of A_τ, the assumed stability, and $\alpha(1+\nu)>1$ complete Assumption C.3. Apply Theorem 3.2. □

Proof of Lemma C.30. Write $\Delta_t = \mathrm{M}_{t+1}-\mathrm{M}_t$, $d_t = \{\mathbb{E}\|\Delta_t\|_\mathrm{F}^2\}^{1/2}$, and $A_k = \sum_{s=1}^k s^{\alpha/2}d_s$. The increment assumption gives $A_k = o(k^{1/2})$. Abel summation applied to $\sum_{s=1}^{t-1} s^{-\alpha/2}(s^{\alpha/2}d_s)$ yields

$$\sum_{s=1}^{t-1} d_s = o\{t^{(1-\alpha)/2}\}. \tag{110}$$

Indeed, its endpoint is $t^{-\alpha/2}o(t^{1/2})$, while its interior sum is bounded by a constant times $\sum_{s<t} o(s^{1/2})s^{-1-\alpha/2}$, which has the same order.

Put $\mathrm{N}_t = \mathrm{M}_t - \mathrm{M}_1$. Minkowski's inequality and (110) give

$$\{\mathbb{E}\|\mathrm{N}_t\|_{\mathrm{F}}^2\}^{1/2} \le \sum_{s=1}^{t-1} d_s = o\{t^{(1-\alpha)/2}\}. \tag{111}$$

The constant-matrix part $n^{-1/2}\sum_t a_t \mathrm{M}_1(e_{t-1}-e_t)$ is $o_{L^1}(1)$ by the summation-by-parts argument in Lemma C.29: Cauchy-Schwarz, $\mathbb{E}\|\mathrm{M}_1\|_{\mathrm{F}}^2 < \infty$, and the parameter rate give the required $O(\eta_t^{1/2})$ bound, including a finite initial term.

It remains to control the N_t part. Because $\mathrm{N}_1 = 0$, discrete summation by parts gives

$$\begin{aligned}\sum_{t=1}^{n} a_t \mathrm{N}_t(e_{t-1}-e_t) =& -a_n \mathrm{N}_n e_n + \sum_{t=1}^{n-1}(a_{t+1}-a_t)\mathrm{N}_{t+1}e_t \\ &+ \sum_{t=1}^{n-1} a_t \Delta_t e_t.\end{aligned}$$

By (111) and the parameter rate, the terminal term divided by $\sqrt{n}$ is $o_{L^1}(1)$. Since $a_{t+1}-a_t = O(t^{\alpha-1})$, the scalar-increment term is bounded in L^1 by

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n-1} t^{\alpha-1} o\{t^{(1-\alpha)/2}\}t^{-\alpha/2} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n-1} o(t^{-1/2}) = o(1).$$

Finally, Cauchy-Schwarz bounds the matrix-increment term by

$$\frac{C}{\sqrt{n}}\sum_{t=1}^{n-1} t^{\alpha/2} d_t \longrightarrow 0.$$

Markov's inequality proves the result. □

Proof of Theorem 3.3. We give the original and weighted arguments in parallel and record separately the three bias terms that enter the weighted recursion.

Step 1: exact original decomposition. Because $\mathrm{M}_t = (\mathrm{P}_t\mathrm{A})^{-1}$, $\mathrm{M}_t\mathrm{P}_t = \mathrm{A}^{-1}$. Write $\Delta\xi_t(e) = \xi_t(\theta_\star + e) - \xi_t(\theta_\star)$. From the original recursion and the mean expansion,

$$\begin{aligned}\eta_t^{-1}\mathrm{M}_t(e_{t-1}-e_t) &= \mathrm{A}^{-1}\widehat{G}_t(\theta_{t-1}) \\ &= e_{t-1} + \mathrm{A}^{-1}\{r(e_{t-1}) + b_t(\theta_\star) + \Delta b_t(\theta_{t-1}) + \psi_t + \Delta\xi_t(e_{t-1})\}.\end{aligned}$$

Therefore

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} e_{t-1} = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t + R_{\mathrm{A}n} - R_{\xi n} - R_{rn} - R_{bn} - R_{b,0,n},$$

where the five remainders correspond, in order, to the Abel term, local innovation, nonlinear mean remainder, local bias, and target bias.

Step 2: Abel and stochastic-equicontinuity terms. The original part of Assumption C.7 and (53) satisfy Lemma C.30; hence $R_{\mathrm{A}n} = o_p(1)$. The local innovation is a martingale difference, so

$$\mathbb{E}\|R_{\xi n}\|^2 \le \frac{C}{n}\sum_{t=1}^{n}\mathbb{E}\|e_{t-1}\|^{\chi} \le \frac{C}{n}\sum_{t=1}^{n}\eta_{t-1}^{\chi/2} \longrightarrow 0.$$

Step 3: nonlinear mean and bias terms. The same calculation as in the fixed-momentum proof gives

$$\mathbb{E}\|R_{rn}\| \le \frac{C}{\sqrt{n}}\sum_{t=1}^{n}\eta_{t-1}^{(1+\nu)/2} \longrightarrow 0.$$

The target bias is negligible by (51). The direct original-path condition (52) gives

$$\mathbb{E}\|R_{bn}\| \le \frac{C}{\sqrt{n}} \sum_{t=1}^{n} \mathbb{E}\|\Delta b_t(\theta_{t-1})\| \longrightarrow 0.$$

Finally, the lagged and ordinary averages differ by $(e_0 - e_n)/\sqrt{n} = o_p(1)$. This proves (7).

Step 4: exact weighted decomposition. Let $\mathcal{F}^\star_{t-1}$ contain the original and replica histories before the current sampling block and multiplier are generated, and write $\Delta\xi^\star_t = \xi_t(\theta^\star_{t-1}) - \xi_t(\theta_\star)$. The exact oracle expansion is

$$\begin{aligned} W_t\widehat{G}_t(\theta^\star_{t-1}) =& \mathrm{A}e^\star_{t-1} + W_t\psi_t + W_t\Delta\xi^\star_t + (W_t - 1)\mathrm{A}e^\star_{t-1} \\ &+ W_t r(e^\star_{t-1}) + W_t\Delta b_t(\theta^\star_{t-1}) + W_t b_t(\theta_\star). \end{aligned}$$

Using $\mathrm{M}^\star_t\mathrm{P}^\star_t = \mathrm{A}^{-1}$ in the weighted recursion and summing therefore gives

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_{t=1}^{n} e^\star_{t-1} =& -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} W_t\psi_t + R^\star_{\mathrm{A}n} \\ &- \mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\big\{W_t\Delta\xi^\star_t + (W_t - 1)\mathrm{A}e^\star_{t-1} + W_t r(e^\star_{t-1}) \\ &\qquad + W_t\Delta b_t(\theta^\star_{t-1}) + W_t b_t(\theta_\star)\big\}, \end{aligned}$$

where $R^\star_{\mathrm{A}n}$ is the dynamic Abel term.

Step 5: weighted remainder bounds. The weighted variation condition and parameter rate give $R^\star_{\mathrm{A}n} = o_{\mathbb{P}\otimes\mathbb{P}^\star}(1)$ by Lemma C.30. The combined array $W_t\Delta\xi^\star_t + (W_t - 1)\mathrm{A}e^\star_{t-1}$ is a martingale-difference array under the enlarged filtration. Since $\mathbb{E}W_t^2 = 2$ and $\mathrm{Var}(W_t) = 1$, martingale orthogonality, (50), and the weighted parameter rate give

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{W_t\Delta\xi^\star_t + (W_t - 1)\mathrm{A}e^\star_{t-1}\}\right\|^2 \le \frac{C}{n}\sum_{t=1}^{n}\{\eta_{t-1}^{\chi/2} + \eta_{t-1}\} \longrightarrow 0.$$

Because $\mathbb{E}|W_t| \le (\mathbb{E}W_t^2)^{1/2}$, the nonlinear mean term is $o_{L^1}(1)$ by $\alpha(1+\nu) > 1$, and (52) gives

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} W_t\Delta b_t(\theta^\star_{t-1})\right\| \le \frac{\mathbb{E}|W_t|}{\sqrt{n}}\sum_{t=1}^{n}\mathbb{E}\|\Delta b_t(\theta^\star_{t-1})\| \longrightarrow 0.$$

For the target bias,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} W_t b_t(\theta_\star) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} b_t(\theta_\star) + \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(W_t - 1)b_t(\theta_\star).$$

The first term is negligible by (51). Conditional on $\mathcal{F}_n$, the second term has mean zero and second moment $n^{-1}\sum_t \|b_t(\theta_\star)\|^2 = o_p(1)$. Thus the target-bias term is $o_{\mathbb{P}\otimes\mathbb{P}^\star}(1)$. Replacing the lagged average costs $(e^\star_0 - e^\star_n)/\sqrt{n} = o_{\mathbb{P}\otimes\mathbb{P}^\star}(1)$ and proves (8).

Step 6: localization and transfer. For local assumptions, apply the preceding steps to the auxiliary paths in (D4), using their uniform effective-inverse variation and rate bounds. Agreement after $k_n = o(\sqrt{n})$ and uniform second moments over the initial segment make the difference between auxiliary and original averages negligible. The globally defined extensions agree with the original oracle at $\theta_\star$, so the target innovation is unchanged. The two linear representations therefore hold for the original paths, and Theorem 3.1 completes the proof. □

Proof of Corollary C.8. For the frozen replica, $\mathrm{P}^\star_t = \mathrm{P}_t$ and therefore $\mathrm{M}^\star_t = \mathrm{M}_t$ at every time. Its initial second moment and (55) follow immediately from the corresponding original-path conditions. The weighted parameter-rate and oracle conditions are assumed in Assumption C.6. Hence Theorem 3.3 applies to the frozen recursion, yielding (8); Theorem 3.1 then gives conditional consistency. □

Proof of Proposition C.9. Because $\mathrm{B}_{t-1}^{\star} \in \mathcal{S}_{\ell,u}$ and metric projection onto a closed convex set is nonexpansive,

$$\|\mathrm{B}_t^{\star} - \mathrm{B}_{t-1}^{\star}\|_{\mathrm{F}} \le \lambda_t \|\widehat{\mathrm{H}}_t^{\star} - \mathrm{B}_{t-1}^{\star}\|_{\mathrm{F}}.$$

Every matrix in $\mathcal{S}_{\ell,u}$ has Frobenius norm at most $u\sqrt{d}$. Hence

$$\{\mathbb{E}\|\mathrm{B}_t^{\star} - \mathrm{B}_{t-1}^{\star}\|_{\mathrm{F}}^2\}^{1/2} \le C\lambda_t \left[1 + \{\mathbb{E}\|\widehat{\mathrm{H}}_t^{\star}\|_{\mathrm{F}}^2\}^{1/2}\right]. \tag{112}$$

Since $\mathrm{P}_{t+1}^{\star} = (\mathrm{B}_t^{\star})^{-1}$,

$$\mathrm{M}_{t+1}^{\star} = (\mathrm{P}_{t+1}^{\star}\mathrm{A})^{-1} = \mathrm{A}^{-1}\mathrm{B}_t^{\star}. \tag{113}$$

Combining the last two displays gives

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{t=1}^{n-1} t^{\alpha/2} \{\mathbb{E}\|\mathrm{M}_{t+1}^{\star} - \mathrm{M}_t^{\star}\|_{\mathrm{F}}^2\}^{1/2} \\
&\quad \le \frac{C}{\sqrt{n}} \sum_{t=1}^{n-1} t^{\alpha/2} \lambda_t \left[1 + \{\mathbb{E}\|\widehat{\mathrm{H}}_t^{\star}\|_{\mathrm{F}}^2\}^{1/2}\right],
\end{aligned}$$

which proves the first claim. Under (58), (112) is $O(t^{-\gamma+\rho/2})$, proving (59). The resulting variation sum is negligible when $\gamma - \rho/2 > (1+\alpha)/2$, equivalently $2\gamma - \rho > 1 + \alpha$. No conditional mean or limit of the curvature observation is used. □

Proof of Theorem 3.4. By Proposition C.9, the joint path satisfies (55). Because $\mathrm{B}_0^{\star} \in \mathcal{S}_{\ell,u}$, (113) at $t = 0$ gives a finite second moment for $\mathrm{M}_1^{\star}$. The remaining hypotheses are precisely those of Theorem 3.3; hence the joint weighted representation (8) holds. Subtracting the original representation and applying Theorem 3.1 proves (9).

Now couple a frozen and a joint replica with the same primary multipliers. Their weighted linear representations have the identical leading term:

$$\begin{aligned}
\sqrt{n}(\bar{\theta}_n^J - \theta_\star) &= -\mathrm{A}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} W_t \psi_t + o_{\mathbb{P}\otimes\mathbb{P}^\star}(1), \\
\sqrt{n}(\bar{\theta}_n^F - \theta_\star) &= -\mathrm{A}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} W_t \psi_t + o_{\mathbb{P}\otimes\mathbb{P}^\star}(1).
\end{aligned}$$

Subtraction gives $\sqrt{n}(\bar{\theta}_n^J - \bar{\theta}_n^F) = o_{\mathbb{P}\otimes\mathbb{P}^\star}(1)$. Applying Lemma C.24 converts this statement to $o_{\mathbb{P}^\star}(1)$ in $\mathbb{P}$-probability, which is (10). □

Proof of Corollary C.10. Put $a = p\kappa$. From (61),

$$\begin{aligned}
\left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} b_t(\theta_\star) \right\| &\le \frac{C}{\sqrt{n}} \sum_{t=1}^{n} t^{-a} = o(1) \qquad (a > 1/2), \\
\frac{1}{n} \sum_{t=1}^{n} \|b_t(\theta_\star)\|^2 &\le \frac{C}{n} \sum_{t=1}^{n} t^{-2a} = o(1).
\end{aligned}$$

For each actual or auxiliary parameter path covered by the statement, the initial-moment bound controls the first summand and (D4) gives

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbb{E}\|\Delta b_t(\vartheta_{t-1})\| \le \frac{C}{\sqrt{n}} + \frac{C}{\sqrt{n}} \sum_{t=2}^{n} t^{-a} \eta_{t-1}^{1/2} \longrightarrow 0.$$

The constants are uniform in the localization index. The limit follows from $a > 1/2$ and $\alpha > 0$; no local squared-bias bound is needed. Hence (51) and (52) hold.

Next,

$$(R_t \delta_t^q)^{-1} \asymp t^{q\kappa_{\mathrm{H}} - r},$$

so (62) implies $\mathbb{E}\|\widehat{\mathrm{H}}_t^{\star}\|_{\mathrm{F}}^2 = O(t^{\rho})$ with $\rho = (q\kappa_{\mathrm{H}} - r)_+$. The second inequality in (63) is therefore exactly (60). Apply Proposition C.9 and Theorem 3.4. □

Proof of Corollary C.11. The proof of Corollary C.5 shows that the direct quantile score has the local mean expansion with derivative A_τ, quadratic-mean continuity exponent $\chi = \nu$, and a finite target second moment. The direct score is exact, so $b_t \equiv 0$ and both bias requirements in (D3) hold automatically. The i.i.d. influence conditions follow from Lemma C.27. The parameter-rate, localization, and original stabilization requirements are assumed in the statement.

From (64),

$$\mathbb{E}\|\widehat{\mathrm{H}}_t^\star\|_{\mathrm{F}}^2 = O(t^\rho), \qquad \rho = (\kappa_{\mathrm{H}} - r)_+.$$

Because $\lambda_t \asymp t^{-1}$, one has $\gamma = 1$, and (65) is equivalent to $2\gamma - \rho > 1 + \alpha$. Therefore Proposition C.9 supplies the joint one-step stabilization. The common-block multiplier rule ensures that $\widehat{\mathrm{H}}_t^\star$ is the weighted version of the original score secant rather than a difference of independently weighted score levels. Apply Corollary C.8 and Theorem 3.4. Finally, A_τ is symmetric, so $\mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-\mathsf{T}} = \mathrm{A}_\tau^{-1}\mathrm{S}_\tau\mathrm{A}_\tau^{-1}$. □

Proof of Proposition C.31. Let $v_t = \mathbb{E}\|e_t\|^2$. Conditional expansion of one update, followed by (89) to (90), gives for all large t,

$$\begin{aligned}\mathbb{E}(\|e_t\|^2 \mid \mathcal{F}_{t-1}) &= \|e_{t-1}\|^2 - 2\eta_t e_{t-1}^{\mathsf{T}}\mathrm{P}_t\mathbb{E}\{\widehat{G}_t(\theta_{t-1}) \mid \mathcal{F}_{t-1}\} \\ &\quad + \eta_t^2\mathbb{E}\{\|\mathrm{P}_t\widehat{G}_t(\theta_{t-1})\|^2 \mid \mathcal{F}_{t-1}\} \\ &\le (1 - 2\mu\eta_t + C\eta_t^2)\|e_{t-1}\|^2 + C\eta_t^2 \\ &\le (1 - \mu\eta_t)\|e_{t-1}\|^2 + C\eta_t^2.\end{aligned}$$

Thus

$$v_t \le (1 - \mu\eta_t)v_{t-1} + C\eta_t^2. \tag{114}$$

Choose $K > C/\mu$ sufficiently large. Since $\eta_{t-1}/\eta_t = (t/(t-1))^\alpha = 1 + O(t^{-1})$ and $t^{-1} = o(\eta_t)$ for $\alpha < 1$, for all sufficiently large t, $(1-\mu\eta_t)K\eta_{t-1} + C\eta_t^2 \le K\eta_t$. After increasing K to dominate the finite initial segment, induction in (114) gives $v_t \le K\eta_t$. For a weighted path, independence and $\mathbb{E}W_t = 1$ preserve the drift term. If the weighted oracle satisfies the stated conditional second-moment bound, the same scalar recursion follows. One sufficient route is the original oracle bound together with $\mathbb{E}W_t^2 = 2$. □

Proof of Lemma C.32. For each coordinate, the martingale version of the von Bahr-Esseen inequality gives $\mathbb{E}|\sum_t D_{tj}|^p \le C_p \sum_t \mathbb{E}|D_{tj}|^p$ (see Hall & Heyde, 1980, Chapter 2). In fixed dimension, $\|x\|^p \le C_d \sum_{j=1}^d |x_j|^p$ and $\sum_j |D_{tj}|^p \le C_d\|D_t\|^p$. Summing the coordinate bounds proves (91). □

Proof of Corollary C.20. Use $g = \nabla R$ and $\psi_t = G(\theta_\star, Z_t)$. Conditions (72) to (74) are exactly the local mean, quadratic-mean continuity, and target second-moment conditions in Assumption C.3. The i.i.d. premise and Lemma C.27 verify Assumption C.1. In the dynamic case, $b_t \equiv 0$, so all bias conditions in (D3) are automatic. The remaining assumptions are the stated algorithmic conditions. Apply Theorem 3.2 in case (a) or Theorem 3.3 in case (b). In either case, Theorem 3.1 gives conditional bootstrap validity. □

Proof of Corollary C.21. For $G(\theta; X, Y) = X(X^{\mathsf{T}}\theta - Y)$ and $e = \theta - \theta_\star$,

$$g(\theta_\star + e) = \mathrm{A}e, \qquad \psi = -X\varepsilon, \qquad \xi(\theta_\star + e) - \xi(\theta_\star) = (XX^{\mathsf{T}} - \mathrm{A})e. \tag{115}$$

Thus the mean remainder and oracle bias are zero, and $\mathbb{E}(\psi\psi^{\mathsf{T}}) = \mathrm{Q}$. The target second moment is finite by (76), so Lemma C.27 applies.

Put $p = 1 + \delta/2$. Since $2p = 2 + \delta$,

$$\mathbb{E}\|XX^{\mathsf{T}} - \mathrm{A}\|_{\mathrm{op}}^p < \infty. \tag{116}$$

For the original path, define $D_t = (X_tX_t^{\mathsf{T}} - \mathrm{A})e_{t-1}$. It is a martingale difference because e_{t-1} is predictable and X_t is independent of the past. The L^2 parameter rate and Lyapunov's inequality give

$$\mathbb{E}\|D_t\|^p \le C\mathbb{E}\|e_{t-1}\|^p \le C\{\mathbb{E}\|e_{t-1}\|^2\}^{p/2} \le C\eta_{t-1}^{p/2}.$$

By Lemma C.32,

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} D_t\right\|^p \le Cn^{-p/2}\sum_{t=1}^{n}\eta_{t-1}^{p/2} = O\{n^{1-p(1+\alpha)/2}\} = o(1), \tag{117}$$

where the last step uses $p(1+\alpha) > 2$.

For a weighted path, the centered local score difference is

$$D_t^\star = W_t(X_tX_t^\mathsf{T} - \mathrm{A})e_{t-1}^\star + (W_t - 1)\mathrm{A}e_{t-1}^\star.$$

It is a martingale difference under the enlarged filtration. Because $p \le 2$, the multiplier assumptions imply $\mathbb{E}|W_t|^p + \mathbb{E}|W_t - 1|^p < \infty$. Using the weighted parameter rate, $\mathbb{E}\|D_t^\star\|^p \le C\eta_{t-1}^{p/2}$, and the same argument as in (117) gives $n^{-1/2}\sum_t D_t^\star = o_{L^p}(1)$.

For dynamic preconditioning, these two estimates replace the quadratic-mean-continuity step in the proof of Theorem 3.3; the Abel and bias terms are unchanged, and the global coefficient bounds avoid auxiliary localization. For fixed momentum, the score along either path has a uniformly bounded pth moment by (115), the parameter rate, and (116). The geometric momentum filter therefore has a uniformly bounded pth moment. In the momentum transformation, this bound is sufficient for the step-size term and gives

$$\mathbb{E}\|e_t - c_\beta\eta_t\mathrm{P}m_t\| \le C\eta_t^{1/2} + C\eta_t,$$

which is the only bound used in the Abel summation. The global proof of Theorem 3.2 then goes through with (117) and its weighted analogue in place of the L^2 local-score calculation. This proves the two asymptotic linear representations and hence the stated bootstrap limit. Under conditional homoskedasticity, $\mathrm{Q} = \mathbb{E}\{\mathbb{E}(\varepsilon^2 \mid X)XX^\mathsf{T}\} = \sigma^2\mathrm{A}$. □

Proof of Corollary C.22. Let $\pi_\theta = \pi(X^\mathsf{T}\theta)$ and $\mathrm{A}(\theta) = \mathbb{E}\{\pi_\theta(1-\pi_\theta)XX^\mathsf{T}\}$. The derivative of $v \mapsto \pi(v)\{1-\pi(v)\}$ is bounded. Hence, with $\nu_\delta = \min(\delta, 1)$ and $1 \wedge x \le x^{\nu_\delta}$,

$$\begin{aligned}\|\mathrm{A}(\theta_\star + e) - \mathrm{A}\|_{\mathrm{op}} &\le C\mathbb{E}\left[\|X\|^2\{1 \wedge (\|X\|\|e\|)\}\right] \\ &\le C\|e\|^{\nu_\delta}\mathbb{E}\|X\|^{2+\nu_\delta}.\end{aligned}$$

These bounds hold for every parameter error. Integration of the derivative of g along the segment from $\theta_\star$ to $\theta_\star + e$ gives $g(\theta_\star + e) = \mathrm{A}e + O(\|e\|^{1+\nu_\delta})$. For the sample score, boundedness and Lipschitz continuity of the logistic link imply

$$\begin{aligned}&\|G(\theta_\star + e; X, Y) - G(\theta_\star; X, Y)\|^2 \\ &\quad\le C\|X\|^2\{1 \wedge (\|X\|^2\|e\|^2)\} \le C\|e\|^\delta\|X\|^{2+\delta}.\end{aligned}$$

Centering changes the bound by at most a constant factor, so quadratic-mean continuity holds with $\chi = \delta$. The target score has a finite second moment because it is bounded in norm by $\|X\|$, and the i.i.d. premise together with Lemma C.27 verifies Assumption C.1. Under the correctly specified Bernoulli model,

$$\begin{aligned}\mathbb{E}(\psi\psi^\mathsf{T} \mid X) &= \mathrm{Var}(Y \mid X)XX^\mathsf{T} \\ &= \pi(X^\mathsf{T}\theta_\star)\{1 - \pi(X^\mathsf{T}\theta_\star)\}XX^\mathsf{T},\end{aligned}$$

so $\mathrm{Q} = \mathrm{A}$. Applying the relevant main theorem gives (79). Under misspecification, the same calculation uses (80) and yields $\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-1}$. □

Proof of Corollary C.23. Let $\zeta_t = U_tU_t^\mathsf{T}s(Z_t)$. Independence of U_t and Z_t and $\mathbb{E}(U_tU_t^\mathsf{T}) = \mathrm{I}_d$ give $\mathbb{E}\zeta_t = 0$ and $\mathbb{E}(\zeta_t\zeta_t^\mathsf{T}) = \mathrm{Q}_U$. By (85), the centered target innovation

$$\psi_t = \widehat{G}_{t,h_t}(\theta_\star) - \mathbb{E}\widehat{G}_{t,h_t}(\theta_\star)$$

satisfies $\mathbb{E}\|\psi_t - \zeta_t\|^2 = O(h_t^4)$. Hence

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\psi_t - \zeta_t)\right\|^2 = O\left(\frac{1}{n}\sum_{t=1}^{n}h_t^4\right) = o(1).$$

The i.i.d. central limit theorem for ζ_t therefore gives $n^{-1/2}\sum_t \psi_t \Longrightarrow N(0, \mathrm{Q}_U)$. The same L^2 approximation, together with truncation, gives empirical covariance convergence and the Lindeberg condition in Assumption C.1; equivalently, one may apply the triangular-array Lindeberg theorem directly.

The bias bounds give

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\|b_t(\theta_\star)\| \le \frac{C}{\sqrt{n}}\sum_{t=1}^{n} t^{-2\kappa} = o(1)$$

when $\kappa > 1/4$, and the square-negligibility condition is immediate. Combining (84) with the $O(\eta_t)$ parameter rate verifies (52). The remaining conditions are assumed, so Theorems 3.1 and 3.3 prove (86) and bootstrap validity.

For $U \sim N(0, \mathrm{I}_d)$ and a fixed vector v, Isserlis' formula gives

$$\mathbb{E}_U\{UU^\mathsf{T} vv^\mathsf{T} UU^\mathsf{T}\} = 2vv^\mathsf{T} + \|v\|^2 \mathrm{I}_d.$$

Taking expectation over $v = s(Z)$ proves (87). □

Proof of Proposition C.12. Conditional on (Z, U), all quantities L_+, L_- and U are fixed. The valid perturbation is

$$(W-1)c_h U, \qquad c_h = \frac{L_+ - L_-}{2h}.$$

Since $\mathrm{Var}(W) = 1$, its covariance is $c_h^2 UU^\mathsf{T}$, which is (66).

For independent side multipliers, set

$$Y_h = \frac{(W^+ - 1)L_+ - (W^- - 1)L_-}{2h} U.$$

The two scalar random variables in the numerator are independent, centered, and have variances L_+^2 and L_-^2. Their cross-covariance is zero. Therefore

$$\mathrm{Cov}^\star(Y_h) = \frac{L_+^2 + L_-^2}{4h^2} UU^\mathsf{T},$$

which is (67).

Taking traces and expectations in the invalid expression gives

$$\frac{1}{4h^2}\mathbb{E}\{(L_+^2 + L_-^2)\|U\|^2\}. \tag{118}$$

Under the stated weighted L^2 convergence, the numerator in (118) converges to

$$2\mathbb{E}\{L(\theta; Z)^2\|U\|^2\} > 0.$$

Hence the trace is asymptotic to a positive constant times h^{-2}.

For the valid expression, suppose there is an L^2 envelope $K(Z, U)$ such that

$$|L(\theta + hU; Z) - L(\theta - hU; Z)| \le 2hK(Z, U), \qquad \mathbb{E}\{K(Z, U)^2\|U\|^2\} < \infty.$$

Then the expected trace in (66) is bounded by $\mathbb{E}\{K(Z, U)^2\|U\|^2\}$, uniformly for small h. □

Proof of Corollary C.13. Replacing L_+ and L_- by $L_+ + C(Z)$ and $L_- + C(Z)$ leaves $L_+ - L_-$ unchanged, so both the finite difference and (66) are unchanged. The numerator of (67), however, becomes

$$\{L_+ + C(Z)\}^2 + \{L_- + C(Z)\}^2 = L_+^2 + L_-^2 + 2C(Z)\{L_+ + L_-\} + 2C(Z)^2.$$

The final two terms are nonzero in general. For a constant shift $C = c$, the quadratic term dominates for sufficiently large $|c|$ whenever $U \neq 0$. Thus the invalid construction depends on an arbitrary loss level that cancels from the original estimator. □

Proof of Proposition C.14. Let $D = \sum_{j=1}^{m} D_j$. With one common multiplier, the perturbation is $(W-1)D$, and therefore

$$\mathrm{Cov}^\star\{(W-1)D\} = DD^\mathsf{T} = \sum_{j=1}^{m}\sum_{k=1}^{m} D_j D_k^\mathsf{T},$$

which proves (68). With independent multipliers, all terms $(W_j - 1)D_j$ are conditionally independent and centered, so their covariances add:

$$\mathrm{Cov}^\star\left\{\sum_{j=1}^{m}(W_j-1)D_j\right\} = \sum_{j=1}^{m} D_j D_j^\mathsf{T}.$$

Subtracting the two expressions leaves precisely the sum of the cross terms with $j \neq k$. □

Proof of Theorem 3.5. Put $X_n = \sqrt{n}(\bar\theta_n - \theta_\star)$. By Theorem 3.1, the conditional law of $T_n^{(b)}$ converges in probability to $N(0, \mathrm{V})$, where $\mathrm{V} = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$. Conditional on the original stream, the B replicas are independent because their multiplier sequences are independent. The original expansion and (35) give $X_n \Longrightarrow N(0, \mathrm{V})$. Applying Lemma C.26 with $Y_{nb} = T_n^{(b)}$ and $Y \sim N(0, \mathrm{V})$ proves (11). The individual weighted remainders may be combined because B is fixed: the probability that their maximum exceeds any fixed threshold is bounded by the sum of B probabilities tending to zero. □

Proof of Corollary C.15. Let $\sigma_a^2 = a^\mathsf{T}\mathrm{V}a > 0$. By Theorem 3.5 and the continuous mapping theorem,

$$\Big(\sqrt{n}\,a^\mathsf{T}(\bar\theta_n - \theta_\star), \sqrt{n}\,a^\mathsf{T}(\bar\theta_n^{(1)} - \bar\theta_n), \ldots\Big) \Longrightarrow \sigma_a(Z_0, Z_1, \ldots, Z_B),$$

where the Z_b are independent standard normal variables. Therefore

$$\frac{a^\mathsf{T}(\bar\theta_n - \theta_\star)}{s_{n,B}(a)} \Longrightarrow \frac{Z_0}{\{B^{-1}\sum_{b=1}^{B} Z_b^2\}^{1/2}} \sim t_B.$$

The zero set of the denominator is closed, whereas the limiting denominator is positive almost surely. The portmanteau theorem therefore gives $\mathbb{P}\{s_{n,B}(a) = 0\} \to 0$. The ratio map is continuous at every point in the support of the limit, and continuity of the t_B distribution gives the stated coverage of (26). □

Proof of Corollary C.16. Write $\mathrm{V_C} = \mathrm{C}\mathrm{V}\mathrm{C}^\mathsf{T}$, which is positive definite. Under Theorem 3.5, after multiplication by $\mathrm{V_C}^{-1/2}$ the original contrast error converges to $Z_0 \sim N_q(0, \mathrm{I}_q)$ and the B replica contrast differences converge to independent $Z_1, \ldots, Z_B \sim N_q(0, \mathrm{I}_q)$, independent of Z_0. Consequently,

$$\Big(\sqrt{n}(\mathrm{C}\bar\theta_n - \mathrm{C}\theta_\star), n\widehat\Sigma_{n,B}(\mathrm{C})\Big) \Longrightarrow \left(\mathrm{V_C}^{1/2} Z_0, \mathrm{V_C}^{1/2}\frac{\mathrm{W}}{B}\mathrm{V_C}^{1/2}\right),$$

where $\mathrm{W} = \sum_{b=1}^{B} Z_b Z_b^\mathsf{T} \sim W_q(\mathrm{I}_q, B)$. Here $W_q(\mathrm{I}_q, B)$ is the Wishart law with scale I_q and B degrees of freedom. Since $B \geq q$, the Gaussian $q \times B$ matrix $(Z_1, \ldots, Z_B)$ has row rank q almost surely, so W is positive definite almost surely.

The singular positive-semidefinite matrices form a closed set. The portmanteau theorem therefore gives

$$\mathbb{P}\{\mathcal{R}_{n,B}(\mathrm{C})^c\} = \mathbb{P}\{\widehat\Sigma_{n,B}(\mathrm{C}) \text{ is singular}\} \longrightarrow 0.$$

Define the same quadratic-form map to be zero on singular matrices. Its discontinuity set is contained in the singular matrices, which the limit avoids almost surely. The continuous mapping theorem therefore gives the displayed limit below for this auxiliary statistic. It differs from $H_{n,B}(\mathrm{C})$ only on $\mathcal{R}_{n,B}(\mathrm{C})^c$, an event with probability tending to zero. Hence

$$H_{n,B}(\mathrm{C}) \Longrightarrow B Z_0^\mathsf{T}\mathrm{W}^{-1}Z_0.$$

The standard normal-Wishart identity gives

$$\frac{B-q+1}{qB}\, BZ_0^{\mathsf{T}}\mathrm{W}^{-1}Z_0 \sim F_{q,B-q+1}.$$

This proves (69). The limiting F law is continuous, so the threshold $c_{q,B}(\zeta)$ gives coverage $1-\zeta$. □

Proof of Corollary C.17. Let

$$Z_n = \sqrt{n}\, a^{\mathsf{T}}(\bar{\theta}_n - \theta_\star), \qquad \sigma_a^2 = a^{\mathsf{T}}\mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}a > 0.$$

By Theorem 3.1, conditionally in probability, $a^{\mathsf{T}}T_n$ converges weakly to $N(0,\sigma_a^2)$, and Z_n converges to the same continuous law. Conditional quantile consistency at $\zeta/2$ and $1-\zeta/2$ therefore gives

$$q_{\zeta/2,n}(a) \xrightarrow{p} \sigma_a z_{\zeta/2},$$
$$q_{1-\zeta/2,n}(a) \xrightarrow{p} \sigma_a z_{1-\zeta/2},$$

where z_p is the standard-normal p-quantile.

The event that the interval (28) contains $a^{\mathsf{T}}\theta_\star$ is exactly

$$q_{\zeta/2,n}(a) \le Z_n \le q_{1-\zeta/2,n}(a).$$

Slutsky's theorem and continuity of the Gaussian distribution give limiting probability $1-\zeta$.

For the implemented interval, conditional on the original stream the B_n replicas are i.i.d. from the law of $a^{\mathsf{T}}T_n$. The conditional Dvoretzky-Kiefer-Wolfowitz inequality gives uniform convergence of their empirical distribution function whenever $B_n \to \infty$. Combined with the continuity and strict increase of the limiting Gaussian distribution at the two target probabilities, the empirical quantiles differ from the ideal conditional quantiles by $o_p(1)$. The same Slutsky argument completes the proof. □

Proof of Proposition C.18. Write $\mathrm{V} = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$. Let $\rho_K : \mathbb{R}^d \to [0,1]$ be continuous, equal to one on $\{\|x\| \le K\}$ and zero on $\{\|x\| \ge K+1\}$. For fixed K, conditional weak convergence in Theorem 3.1 gives

$$\mathbb{E}^\star\{T_n\rho_K(T_n)\} \xrightarrow{p} \mathbb{E}\{Z\rho_K(Z)\},$$
$$\mathbb{E}^\star\{T_nT_n^{\mathsf{T}}\rho_K(T_n)\} \xrightarrow{p} \mathbb{E}\{ZZ^{\mathsf{T}}\rho_K(Z)\}, \qquad Z \sim N(0,\mathrm{V}).$$

Conditional independence and boundedness of the truncated variables give conditional laws of large numbers for these moments at each fixed K.

To control the omitted tails, put $q_{n,K} = \mathbb{E}^\star\{\|T_n\|^2\mathbf{1}(\|T_n\| > K)\}$, allowing the value $+\infty$. For $\varepsilon,\delta > 0$, conditional Markov's inequality on $\{q_{n,K} \le \delta\}$ gives

$$(\mathbb{P}\otimes\mathbb{P}^\star)\left\{\frac{1}{B_n}\sum_{b=1}^{B_n}\|T_n^{(b)}\|^2\mathbf{1}(\|T_n^{(b)}\| > K) > \varepsilon\right\} \le \mathbb{P}(q_{n,K} > \delta) + \frac{\delta}{\varepsilon}.$$

Taking $n\to\infty$, then $K\to\infty$, and then $\delta\downarrow 0$ makes this bound vanish by (70). First-moment tails are bounded using $\|x\|\mathbf{1}(\|x\| > K) \le K^{-1}\|x\|^2\mathbf{1}(\|x\| > K)$. The corresponding Gaussian tails vanish because Z has finite second moment. Consequently,

$$\frac{1}{B_n}\sum_{b=1}^{B_n} T_n^{(b)} \xrightarrow{p} 0, \qquad \frac{1}{B_n}\sum_{b=1}^{B_n} T_n^{(b)}\{T_n^{(b)}\}^{\mathsf{T}} \xrightarrow{p} \mathrm{V}.$$

Subtracting the outer product of the empirical mean and multiplying by $B_n/(B_n-1) \to 1$ proves (71). The argument uses only finite truncated conditional moments and does not require an ideal conditional covariance to exist on every original-data event.

□

Proof of Corollary C.19. Write $\mathrm{V} = \mathrm{A}^{-1}\mathrm{Q}\mathrm{A}^{-\mathsf{T}}$, let C be the $q \times d$ matrix whose jth row is a_j^{T}, and put $\mathrm{D} = \mathrm{diag}(\mathrm{CVC}^{\mathsf{T}})^{1/2}$. Positivity of the contrast variances makes D nonsingular. Let

$$Z_{\max} = \|\mathrm{D}^{-1}\mathrm{C}Z\|_\infty, \qquad Z \sim N(0, \mathrm{V}),$$

and denote its $(1-\zeta)$ quantile by $c_{1-\zeta}$. The distribution of $Z_{\max}$ is continuous because each coordinate of $\mathrm{D}^{-1}\mathrm{C}Z$ is nondegenerate and q is fixed. Its distribution function is strictly increasing on $(0, \infty)$: let L be the linear support of $\mathrm{D}^{-1}\mathrm{C}Z$. For $0 < a < b$, scaling any nonzero vector in L shows that $\{y \in L : a < \|y\|_\infty < b\}$ is nonempty and relatively open. The Gaussian law has a positive density on L, so this set has positive probability. Thus the $(1-\zeta)$ quantile is unique, even if the joint contrast covariance is singular.

For a generic replica, define the fixed-denominator statistic

$$\widetilde{S}_n = \max_{1 \le j \le q} \frac{|a_j^{\mathsf{T}} T_n|}{(a_j^{\mathsf{T}} \mathrm{V} a_j)^{1/2}}.$$

By Theorem 3.1 and the continuous mapping theorem, its conditional law converges in probability to the law of $Z_{\max}$. Conditional on the original stream, the replicas are i.i.d.; hence their empirical distribution at every continuity point converges to the conditional distribution of $\widetilde{S}_n$ as $B_n \to \infty$.

It remains to replace V by the common empirical covariance $\widehat{\mathrm{V}}_{n,B_n}$, which is computed from the same replicas. By Proposition C.18, all q estimated contrast variances converge to their positive limits. Hence the fallback event on which the output is empty has probability tending to zero. On the event $\|T_n^{(b)}\| \le K$, the resulting change in each studentized statistic is bounded uniformly in b by K times the largest relative contrast-variance error, and is therefore $o_p(1)$ for fixed K. The fraction of replicas outside this event is, conditionally, asymptotically no larger than the conditional tail probability of $\|T_n\|$; tightness from Theorem 3.1 makes this arbitrarily small by taking K large. Thus the empirical distribution of the actually computed statistics in (31) converges, at every continuity point, to the distribution of $Z_{\max}$. Consequently,

$$\widehat{c}_{1-\zeta,n,B_n} \xrightarrow{p} c_{1-\zeta}. \tag{119}$$

For the original estimator, Theorem 3.1 and Proposition C.18 and Slutsky's theorem give

$$\max_{1 \le j \le q} \frac{|\sqrt{n}\, a_j^{\mathsf{T}}(\bar{\theta}_n - \theta_\star)|}{(a_j^{\mathsf{T}} \widehat{\mathrm{V}}_{n,B_n} a_j)^{1/2}} \Longrightarrow Z_{\max}.$$

Combining this display with (119) shows that the probability that all inequalities defining (32) hold converges to $1-\zeta$. □

Proof of Proposition C.33. The proof of Theorem 3.3, retaining the target-bias sum instead of imposing the first limit in (51), gives

$$\sqrt{n}(\bar{\theta}_n - \theta_\star) = -\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t - \mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} b_t(\theta_\star) + o_p(1). \tag{120}$$

Because $h_t^p = h_0^p t^{-1/2}$ and $n^{-1/2}\sum_{t=1}^{n} t^{-1/2} \to 2$,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} b_t(\theta_\star) \xrightarrow{p} 2h_0^p b_{\mathrm{lead}}.$$

Combining this with the influence central limit theorem proves (93).

For the centered weighted statistic, the bias contribution after subtraction is

$$-\mathrm{A}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(W_t - 1) b_t(\theta_\star).$$

Its conditional second moment is bounded by

$$\frac{C}{n}\sum_{t=1}^{n}\|b_t(\theta_\star)\|^2 = O\left(\frac{\log n}{n}\right) + o_p(1),$$

where the remainder follows from (92) and the retained second limit in (51). The bound converges to zero in probability, so conditional Chebyshev makes the term conditionally negligible. This proves the mismatch.

Adding (95) to (120) cancels the bias sum by (94): the residual correction is the sum of $\widehat{\mathrm{A}}^{-1} n^{-1/2} \sum_t \{\widehat{b}_t - b_t(\theta_\star)\} = o_p(1)$ and $(\widehat{\mathrm{A}}^{-1} - \mathrm{A}^{-1}) n^{-1/2} \sum_t b_t(\theta_\star) = o_p(1)$, since the latter bias sum converges to a fixed vector. Slutsky's theorem gives the centered limit. Finally, substituting (96) into (97) gives

$$\begin{aligned}\mathbb{E}\{\widetilde{G}_{t,h}(\theta) \mid \mathcal{F}_{t-1}\} &= g(\theta) + \frac{2^p c_t(\theta)(h/2)^p - c_t(\theta)h^p}{2^p - 1} + O(h^{p+\delta_b}) \\ &= g(\theta) + O(h^{p+\delta_b}).\end{aligned}$$

The multiplier statement follows from the weighting rule in Section 2.2. □