INFINITE QUANTILES INFERENCE UNDER LOCAL DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study simultaneous quantile inference under local differential privacy for a number of target probabilities that grows with the sample size. Each user participates once, so learning unknown quantile locations and calibrating their joint uncertainty must share the same private information. We fuse complementary cumulative distribution function (CDF) scores from a single tree report using their exact covariance; the variance at the correct anchor is no larger than that of either score, and anchor error incurs only quadratic excess variance. Three consecutive cohorts then localize the quantiles, estimate their densities, and supply fresh scores for a one-step correction and joint multiplier calibration. Under local density regularity, uniform identification on a public bracket, and explicit growth conditions, a uniform representation controls localization and density errors and yields asymptotically valid simultaneous coverage. A fixed-privacy regime permits every polynomially growing target grid. Simulations show consistent same-report precision gains, while comparisons with CDF-band inversion expose the sample cost of learning target locations. Census microdata illustrates how the resulting joint intervals support wage-percentile ratios.

1 INTRODUCTION

From 1987 to 2004, U.S. upper-half wage inequality rose while lower-half inequality contracted (Autor et al., 2006). Quantile comparisons distinguish these distributional changes that a mean or an isolated percentile can miss. As a data stream grows, a finer percentile grid permits more detailed comparisons. Their uncertainty must be assessed jointly: a collection of pointwise intervals does not control the largest error over an expanding grid. We study this problem when each user privatizes a measurement before sending it to the server and participates only once.

Local differential privacy (LDP) makes this joint inference problem different from classical quantile inference. Sample-quantile representations and cumulative distribution function (CDF) bands connect nonprivate quantile error to a shared empirical process (Bahadur, 1966; Csörgő, 1983; Massart, 1990; Chernozhukov et al., 2013b). Under LDP, the server receives randomized reports rather than observations (Warner, 1965; Duchi et al., 2018). It must learn where the quantiles are while deciding which thresholds to query. Giving each percentile a separate user group loses information as the number of targets grows; using a shared report avoids this split but creates dependent errors at adaptively selected thresholds. Joint calibration must accommodate both that dependence and the error in the learned locations.

Recent work provides online inference for a prespecified scalar quantile. Liu et al. (2023) use randomized binary inquiries and self-normalization to avoid nuisance-density estimation. Cai et al. (2025) use dynamically growing parallel stochastic-approximation paths to obtain an asymptotic confidence sequence over time. Other developments address communication-efficient single-quantile estimation, federated scalar inference, and fixed-dimensional quantile-regression contrasts (Aamand et al., 2025; Cai et al., 2026; Liu and Hu, 2026). Our target is joint inference at a terminal horizon fixed before data collection, with the number of probability targets increasing as that horizon grows. Extending scalar inference to this setting requires uniform control of localization and density errors at the scale of a maximum, rather than a separate scalar approximation at every target.

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.

CDF estimation followed by inversion is an important alternative. Liu et al. (2024) study private CDF estimation on public inspection grids. A simultaneous CDF band supports probability queries selected after release, whereas direct inference learns locations for prespecified probabilities.

For independent observations with CDF F, our targets are

$$0 < \tau_{1,n} < \cdots < \tau_{k_n,n} < 1, \qquad Q_{j,n} = F^{-1}(\tau_{j,n}),$$

where the deterministic number k_n may diverge. "Infinite" refers to this increasing sequence of finite target grids. With one report per user at privacy budget $\varepsilon_n > 0$ and public queries allowed to depend on past reports, we seek intervals satisfying

$$\mathbb{P}\{Q_{j,n} \in \mathcal{I}_{n,j} \text{ for every } j \le k_n\} \longrightarrow 1 - \alpha, \qquad \alpha \in (0, 1).$$

The probability targets are specified before collection; the method learns their spatial locations.

Our first step extracts more information from a shared range-tree report with optimized unary encoding (OUE) (Wang et al., 2017; Cormode et al., 2019). Every threshold has two unbiased CDF scores: a prefix sum and the complement of a suffix sum. Their exact covariance determines a minimum-variance linear fusion at the correct CDF anchor. Fusion remains unbiased for any public anchor, and the variance penalty for an inaccurate anchor is quadratic. Thus the target probability itself supplies an increasingly accurate anchor as localization improves. This gain changes only decoding, with no additional report or privacy expenditure.

Our second step separates adaptive learning from joint calibration. The first cohort averages one projected stochastic-approximation path for each target. The second estimates densities near the learned locations. The third supplies fresh user-level score vectors. A one-step correction cancels first-order localization error, leaving a product of localization and density errors and a local nonlinear remainder. Conditional on the first two cohorts, the fresh vectors are independent and identically distributed; shared Gaussian multipliers retain their cross-quantile dependence. We bound the remainder uniformly and connect it to high-dimensional Gaussian approximation (Polyak and Juditsky, 1992; Chernozhukov et al., 2013a; 2017). At fixed privacy, a sufficient regime permits every polynomial growth rate of k_n, with only logarithmic grid-size factors in the width bound.

We compare a same-channel fixed-grid inversion procedure with the complete three-cohort method, including the users spent on learning. Fusion reduces error by 21–23% relative to balanced decoding in the matched three-stage experiments. CDF inversion gives smaller errors and narrower, more conservative intervals on the tested public spatial grids. These comparisons distinguish the benefit of extracting complementary information from a report from the cost of learning where to query it. The paper is organized as follows. Section 2 develops the procedure, and Section 3 establishes its guarantees. Sections 4 and 5 present simulations and wage-percentile comparisons; Section 6 concludes.

2 Shared private scores and quantile inference

Algorithm 1 gives the complete procedure, from public inputs to simultaneous intervals. We first construct the shared private report and then describe the three cohorts; Appendix A.2 gives the streaming updates, and Appendix A.3 specifies the default tuning.

We suppress the subscript n on $\tau_{j,n}$ and $Q_{j,n}$ when the sample size is clear. The generalized inverse is

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \ge u\}.$$

The public projection bracket $[L, U]$ contains all target quantiles in its interior, while the raw observations may be unbounded. At time i, a public query q_i may be a measurable function of the previous reports $Z_1, \ldots, Z_{i-1}$. A local randomizer $\mathcal{M}_i$ is ε_n-LDP conditional on the public history if, for every pair x, x', every history, and every measurable report set A,

$$\mathbb{P}\{\mathcal{M}_i(x; \mathsf{q}_i) \in A \mid Z_{1:i-1}\} \le e^{\varepsilon_n} \mathbb{P}\{\mathcal{M}_i(x'; \mathsf{q}_i) \in A \mid Z_{1:i-1}\}.$$

The complete procedure is user-level ε_n-LDP for every deterministic horizon. Conditional on the past transcript, OUE has this privacy bound and the sampled tree level is independent of the user's value. History-dependent queries and server-side computations preserve the bound; each person belongs to only one cohort. Appendix G gives the likelihood-ratio argument for the full sequential transcript.

2.1 COMPLEMENTARY SCORES FROM ONE REPORT

For privacy budget $\varepsilon = \varepsilon_n$, fix d ordered public thresholds

$$x_{(1)} \le \cdots \le x_{(d)}.$$

Stable tags distinguish ties. These thresholds induce $d+1$ interval categories

$$C_0 = (-\infty, x_{(1)}], \quad C_s = (x_{(s)}, x_{(s+1)}]\ (1 \le s < d), \quad C_d = (x_{(d)}, \infty).$$

Let $H_d = \lceil \log_2(d+1) \rceil$ and pad the $d+1$ real leaves by zero-probability leaves to obtain a complete binary tree with 2^{H_d} leaves. Levels are numbered $1, \ldots, H_d$, with the root excluded. A user with value X determines the leaf $C(X)$, draws a level $L^\star$ uniformly from $\{1, \ldots, H_d\}$, and finds the unique node on the path to $C(X)$ at that level (Cormode et al., 2019). The node is represented by a unary vector. With $p_\varepsilon = 1/2, q_\varepsilon = 1/(e^\varepsilon + 1)$, OUE (Wang et al., 2017) independently releases every bit as one with probability p_ε when its input bit is one and with probability q_ε when its input bit is zero. The report consists of the sampled level and the privatized unary vector.

For a node v at level $\ell(v)$, let Z_v be its released bit, or zero on an unsampled level, and define

$$\psi_v(Z) = H_d \mathbf{1}\{L^\star = \ell(v)\} \frac{Z_v - q_\varepsilon}{p_\varepsilon - q_\varepsilon}.$$

The privacy-noise constant and a deterministic score envelope are

$$\nu_\varepsilon = \frac{4e^\varepsilon}{(e^\varepsilon - 1)^2}, \qquad \mathcal{B}_{d,\varepsilon} = 1 + \frac{2H_d e^\varepsilon}{e^\varepsilon - 1}. \tag{2.1}$$

For threshold rank $r \in \{1, \ldots, d\}$, let $\mathcal{D}_r^-$ be the canonical disjoint-node decomposition of the first r leaves and let $\mathcal{D}_r^+$ be the decomposition of the complementary leaves. Each decomposition consists of the maximal tree nodes whose leaves lie entirely in the required set. Write $w_r^- = |\mathcal{D}_r^-|$ and $w_r^+ = |\mathcal{D}_r^+|$; both are between one and H_d, and each decomposition contains at most one node at any level. Define

$$A_r = \sum_{v \in \mathcal{D}_r^-} \psi_v, \qquad B_r = 1 - \sum_{v \in \mathcal{D}_r^+} \psi_v.$$

Both estimate $F(x_{(r)})$. We choose the fusion weight by evaluating the variances and covariance of A_r and B_r at a public reference value $u \in (0,1)$ for $F(x_{(r)})$. For this anchor, set

$$\begin{aligned} a_r^-(u) &= H_d w_r^- \nu_\varepsilon + (2H_d - 1)u, \\ a_r^+(u) &= H_d w_r^+ \nu_\varepsilon + (2H_d - 1)(1-u), \\ \lambda_r^\circ(u) &= \frac{a_r^+(u)}{a_r^-(u) + a_r^+(u)}, \end{aligned}$$

and form the fused score

$$G_r(u) = \lambda_r^\circ(u) A_r + \{1 - \lambda_r^\circ(u)\} B_r. \tag{2.2}$$

Proposition 2.1 (Exact fused-score moments). *Let $p_r = F(x_{(r)})$. Condition on the pre-report public history, including the threshold list and the anchor u, so both are fixed. Then*

$$\begin{aligned} \mathbb{E}A_r = \mathbb{E}B_r &= p_r, \\ \mathrm{Var}(A_r) &= H_d w_r^- \nu_\varepsilon + 2H_d p_r - p_r^2, \\ \mathrm{Var}(B_r) &= H_d w_r^+ \nu_\varepsilon + 2H_d(1-p_r) - (1-p_r)^2, \\ \mathrm{Cov}(A_r, B_r) &= p_r(1-p_r). \end{aligned}$$

Consequently, $G_r(u)$ is unbiased for every anchor fixed by this history. At $u = p_r$,

$$\mathrm{Var}\{G_r(p_r)\} = p_r(1-p_r) + \frac{a_r^-(p_r) a_r^+(p_r)}{a_r^-(p_r) + a_r^+(p_r)} \le \min\{\mathrm{Var}(A_r), \mathrm{Var}(B_r)\}. \tag{2.3}$$

For an arbitrary anchor u,

$$\mathrm{Var}\{G_r(u)\} - \mathrm{Var}\{G_r(p_r)\} = \{a_r^-(p_r) + a_r^+(p_r)\}\{\lambda_r^\circ(u) - \lambda_r^\circ(p_r)\}^2, \tag{2.4}$$

with the explicit bound

$$0 \le \mathrm{Var}\{G_r(u)\} - \mathrm{Var}\{G_r(p_r)\} \le (2H_d - 1)(u - p_r)^2. \tag{2.5}$$

Moreover, $|G_r(u)| \le \mathcal{B}_{d,\varepsilon}$ almost surely.

At the correct anchor, fusion minimizes variance among linear combinations with weights summing to one. The shorter-side balanced decoder instead chooses A_r or B_r by node count. All stages use $u = \tau_j$ for target j, including both density-estimation thresholds. Unbiasedness holds throughout adaptation; (2.5) makes the variance cost vanish quadratically as the queried CDF value approaches τ_j. The proof is in Appendix F.

2.2 Adaptive localization

The total horizon n and stage fractions are fixed before data collection. Set

$$n_1 = \lfloor \rho_1 n \rfloor, \qquad n_2 = \lfloor \rho_2 n \rfloor, \qquad n_3 = n - n_1 - n_2, \tag{2.6}$$

where $\rho_s > 0$ and $\rho_1 + \rho_2 + \rho_3 = 1$. Our default is $(\rho_1, \rho_2, \rho_3) = (0.35, 0.30, 0.35)$. Each person contributes to exactly one stage; the final score vectors are fresh relative to both learning stages.

Initialize deterministic $q_{0,j} \in [L, U]$. Choose $a \in (1/2, 1)$ and $\gamma_t = \gamma_0 t^{-a}$ with $\gamma_0 > 0$. At Stage-1 arrival t, the server sorts the current thresholds $q_{t-1,1}, \ldots, q_{t-1,k_n}$, retains their target labels, and sends the ordered list to the user. The user releases one tree-OUE report. If target j has current rank $r_t(j)$, the server decodes using (2.2)

$$G_{t,j}^{(1)} = G_{r_t(j)}(\tau_j)$$

after which it updates

$$q_{t,j} = \Pi_{[L,U]} \left[q_{t-1,j} - \gamma_t \{ G_{t,j}^{(1)} - \tau_j \} \right]. \tag{2.7}$$

Here $\Pi_{[L,U]}$ clips an iterate to the public bracket. The score is conditionally unbiased:

$$\mathbb{E}\{ G_{t,j}^{(1)} \mid \mathcal{F}_{t-1} \} = F(q_{t-1,j}),$$

where $\mathcal{F}_{t-1}$ contains the previous private transcript. Stable tags keep each target attached to its CDF score through coordinate crossings.

The preliminary center is the full Polyak–Ruppert average

$$q_j^0 = \overline{q}_{n_1,j} = \frac{1}{n_1} \sum_{t=1}^{n_1} q_{t,j}. \tag{2.8}$$

Both the iterates and their average are updated in $O(k_n)$ memory. All iterates enter the average; the online identity and tail-average comparison are given in Appendix A.2.

2.3 Density learning and fresh-score calibration

Stage 2 freezes the preliminary centers and queries $q_j^0 \pm h_n$ with deterministic bandwidth $h_n > 0$. The $2k_n$ tagged thresholds are sorted once. Every Stage-2 user sends one tree-OUE report, from which the server obtains fused scores $G_{i,j,-}^{(2)}$ and $G_{i,j,+}^{(2)}$. Let

$$\overline{G}_{j,\pm}^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} G_{i,j,\pm}^{(2)}$$

and define

$$\widetilde{f}_j = \frac{\overline{G}_{j,+}^{(2)} - \overline{G}_{j,-}^{(2)}}{2h_n}. \tag{2.9}$$

For a deterministic sequence $\varpi_n \downarrow 0$, use the vanishing stabilization

$$\widehat{f}_j = \min\{ \varpi_n^{-1}, \max(\varpi_n, \widetilde{f}_j) \}. \tag{2.10}$$

Consistency makes the clipping inactive with probability tending to one.

Stage 3 again freezes the tagged list $q_1^0, \ldots, q_{k_n}^0$. From each new user's single report, the server obtains a vector

$$G_i^{(3)} = (G_{i,1}^{(3)}, \ldots, G_{i,k_n}^{(3)})^\top, \qquad \overline{G}_j^{(3)} = n_3^{-1} \sum_{i=1}^{n_3} G_{i,j}^{(3)}.$$

The one-step estimator is

$$\check{Q}_j = q_j^0 + \frac{\tau_j - \overline{G}_j^{(3)}}{\widehat{f}_j}. \tag{2.11}$$

Conditional on Stages 1 and 2, the Stage-3 vectors are independent and identically distributed. Write

$$\widehat{\sigma}_{G,j}^2 = \frac{1}{n_3}\sum_{i=1}^{n_3}(G_{i,j}^{(3)} - \overline{G}_j^{(3)})^2.$$

Write $\widehat{\mathcal{J}}_n = \{j : \widehat{\sigma}_{G,j} > 0\}$; a zero-variance coordinate receives $[L, U]$. Take an empty maximum as zero. For each draw b, one independent $\xi_i^{(b)} \sim N(0,1)$ multiplies all coordinates of user i's centered score vector. Sharing this multiplier across coordinates preserves their empirical correlation in the simulated vector. For $1 \le b \le M$, define

$$T_b^* = \max_{j \in \widehat{\mathcal{J}}_n} \left| \frac{n_3^{-1/2}\sum_{i=1}^{n_3}\xi_i^{(b)}(G_{i,j}^{(3)} - \overline{G}_j^{(3)})}{\widehat{\sigma}_{G,j}} \right|. \tag{2.12}$$

Let $T_{(1)}^* \le \cdots \le T_{(M)}^*$ be the ordered multiplier statistics and set

$$r_{M,\alpha} = \min\{M, \lceil (M+1)(1-\alpha) \rceil\}, \qquad \widehat{c}_{1-\alpha}^* = T_{(r_{M,\alpha})}^*. \tag{2.13}$$

The simultaneous intervals are

$$\mathcal{I}_{n,j} = \left[\check{Q}_j - \frac{\widehat{c}_{1-\alpha}^*\widehat{\sigma}_{G,j}}{\widehat{f}_j\sqrt{n_3}}, \check{Q}_j + \frac{\widehat{c}_{1-\alpha}^*\widehat{\sigma}_{G,j}}{\widehat{f}_j\sqrt{n_3}} \right] \cap [L, U]. \tag{2.14}$$

The same reporting rule serves all three stages, with k_n thresholds for localization and inference and $2k_n$ for density estimation. Its unary message uses $O(k_n)$ bits. Maintaining running score sums and multiplier-weighted sums gives $O(Mk_n)$ server memory, independent of the horizon; Appendix A.2 gives the identities.

3 Theoretical Results

Assume i.i.d. observations, target probabilities in a fixed interior subinterval of $(0,1)$, and target quantiles uniformly inside the public bracket. On fixed neighborhoods of these quantiles, the density satisfies $0 < f_- \le f \le f_+ < \infty$ and is β-Hölder, $0 < \beta \le 1$. On $[L, U]$, uniform identification means $\mu|q - Q_j| \le |F(q) - \tau_j| \le L_F|q - Q_j|$ for fixed $0 < \mu \le L_F < \infty$. The gain constant satisfies $0 < \gamma_0 \le (2L_F)^{-1}$; Assumption D.4 gives the localization conditions. The sequence b_n there is an analytical localization index; it does not delete observations from the full average.

Write $\ell_n = \log\{2k_n(n+1)\}$, $B_{1n} = \mathcal{B}_{k_n,\varepsilon_n}$ and $B_{2n} = \mathcal{B}_{2k_n,\varepsilon_n}$. The localization and density rates are

$$r_{1n} = b_n/n_1 + B_{1n}\sqrt{\ell_n/n_1} + B_{1n}n_1^{a/2-1}\sqrt{\ell_n} + B_{1n}^{1+\beta}n_1^{-a(1+\beta)/2}\ell_n^{(1+\beta)/2},$$
$$d_n = r_{1n}^\beta + h_n^\beta + (B_{2n}/h_n)\{\sqrt{\ell_n/n_2} + \ell_n/n_2\}.$$

Here $h_n \to 0$, $d_n \to 0$, and $\varpi_n \downarrow 0$. Conditional on the first two cohorts, define $Y_{i,j} = G_{i,j}^{(3)} - F(q_j^0)$, $\sigma_{j,n}^2 = \mathbb{E}(Y_{i,j}^2 \mid \mathcal{F}_{n_1+n_2})$, $\underline{\sigma}_n = \min_j \sigma_{j,n}$ and $K_n = 2B_{1n}/\underline{\sigma}_n$. Gaussian approximation and transfer of the one-step remainder require

$$K_n^2\ell_n^7/n_3 = o_{\mathbb{P}}(1), \qquad \sqrt{n_3\ell_n}\{r_{1n}d_n + r_{1n}^{1+\beta}\}/\underline{\sigma}_n = o_{\mathbb{P}}(1).$$

These conditions make both the score approximation and the accumulated localization error negligible on the simultaneous scale. Appendix D.1 states the full assumptions; Corollary 3.3 below gives a directly interpretable sufficient growth regime.

The Stage-1 and Stage-2 arguments are collected in Appendices H and I. They establish the uniform rates $\max_j |q_j^0 - Q_{j,n}| = O_{\mathbb{P}}(r_{1n})$ and $\max_j |\widehat{f}_j - f(Q_{j,n})| = O_{\mathbb{P}}(d_n)$; see Theorems H.4 and I.2. The term r_{1n}^β in d_n accounts for centering the finite difference at the estimated quantile. For $\beta = 1$, balancing its bandwidth and stochastic terms gives $h_n \asymp \{B_{2n}^2\ell_n/n_2\}^{1/4}$.

Theorem 3.1 (Uniform one-step representation)**.** *Suppose Assumptions D.1, D.2 and D.5 hold and* (H.23) *and* (I.1) *are valid. Then*

$$\max_{j\le k_n}\left|\widehat{f}_j(\check{Q}_j - Q_{j,n}) + \frac{1}{n_3}\sum_{i=1}^{n_3} Y_{i,j}\right| = O_{\mathbb{P}}(R_n), \qquad R_n = r_{1n}d_n + r_{1n}^{1+\beta}. \tag{3.1}$$

The conditional mean of $\tau_j - \overline{G}_j^{(3)}$ is $\tau_j - F(q_j^0)$. Dividing by $\widehat{f}_j$ therefore cancels the linear location error. The remainder R_n has a direct interpretation: estimating the derivative costs $r_{1n}d_n$, and linearizing F costs $r_{1n}^{1+\beta}$. The simultaneous transfer condition above makes these errors small even after taking the maximum over targets.

Let $c^*_{1-\alpha}$ denote the exact conditional $(1-\alpha)$ quantile of (2.12) given the observed Stage-3 scores and the first two stages. Let $\mathcal{I}^*_{n,j}$ denote the interval in (2.14) with $\widehat{c}^*_{1-\alpha}$ replaced by $c^*_{1-\alpha}$.

Theorem 3.2 (Growing-grid multiplier validity)**.** *Suppose Assumptions D.1 to D.6 hold. Define*

$$T_n = \max_{j\in\widehat{\mathcal{J}}_n}\left|\frac{\sqrt{n_3}\widehat{f}_j(\check{Q}_j - Q_{j,n})}{\widehat{\sigma}_{G,j}}\right|.$$

Then, for every fixed $0 < \alpha_0 < 1/2$*,*

$$\sup_{\alpha\in[\alpha_0,1-\alpha_0]}\left|\mathbb{P}\{T_n \le c^*_{1-\alpha}\} - (1-\alpha)\right| \longrightarrow 0. \tag{3.2}$$

Consequently,

$$\mathbb{P}\{Q_{j,n} \in \mathcal{I}^*_{n,j} \text{ for all } j \le k_n\} \longrightarrow 1-\alpha. \tag{3.3}$$

If the empirical quantile based on $M_n \to \infty$ *multiplier draws is used, then* (3.2) *and* (3.3) *remain valid with* $\widehat{c}^*_{1-\alpha}$ *and* $\mathcal{I}_{n,j}$ *in place of* $c^*_{1-\alpha}$ *and* $\mathcal{I}^*_{n,j}$*, respectively. Here the probability also averages over the independent calibration multipliers. If instead* M *is fixed, then for every fixed* $\alpha \in (0,1)$*,*

$$\mathbb{P}\{T_n \le \widehat{c}^*_{1-\alpha}\} \longrightarrow \frac{r_{M,\alpha}}{M+1}, \tag{3.4}$$

and the simultaneous coverage probability has the same limit.

The multiplier uses the dependence among neighboring quantiles through each user's score vector, calibrating their maximum error jointly.

Corollary 3.3 (Width and grid growth)**.** *Under the assumptions of Theorem 3.2,*

$$\max_{j\le k_n}\mathrm{halfwidth}(\mathcal{I}_{n,j}) = O_{\mathbb{P}}\left\{\frac{H_{k_n}\sqrt{1+\nu_{\varepsilon_n}}\sqrt{\log(2k_n)}}{f_-\sqrt{n_3}}\right\}. \tag{3.5}$$

For $0 < \varepsilon_n \le 1$*, this is*

$$O_{\mathbb{P}}\left\{\frac{H_{k_n}\sqrt{\log(2k_n)}}{\varepsilon_n f_-\sqrt{n_3}}\right\}. \tag{3.6}$$

If ε_n *is bounded away from zero,* $n_s/n \to \rho_s \in (0,1)$*,* $a = 2/3$*,* $\beta = 1$*, and*

$$\frac{\log^9\{2k_n(n+1)\}}{n} \longrightarrow 0, \tag{3.7}$$

then the conditions of Theorem 3.2 hold for the bandwidth

$$h_n \asymp \left\{\frac{B_{2n}^2\ell_n}{n_2}\right\}^{1/4}$$

provided $b_n/n = o(\ell_n^{3/2}n^{-1/2})$*. In particular, every polynomially growing* k_n *is allowed.*

The width bound in (3.5) separates the private-score scale $H_{k_n}\sqrt{1+\nu_{\varepsilon_n}}$ from the simultaneous calibration factor $\sqrt{\log(2k_n)}$. The density lower bound converts CDF error into quantile error. Thus, at fixed privacy and density bounds, grid size enters this upper bound only through logarithmic factors, while precision improves at the usual square-root rate in the inference cohort size n_3. Condition (3.7) combines the tree-score envelope, density estimation, Gaussian approximation, and anti-concentration in one simple rate. It allows $k_n = \exp\{o(n^{1/9})\}$, including every polynomially growing grid.

4 NUMERICAL EXPERIMENTS

We distinguish precision gained by decoding the same private report from the performance of complete inferential procedures. Write $k = k_n$. Each setting uses 500 independent Monte Carlo repetitions, 500 Gaussian multiplier draws, and nominal simultaneous coverage 0.95. Probabilities are equally spaced on $[0.05, 0.95]$. Coverage records whether the entire grid is covered in a repetition; at 95%, its Monte Carlo standard error is about one percentage point. For point estimates $\widehat{Q}_{r,j}$ in repetition r, the pooled root mean squared error is

$$\mathrm{RMSE} = \left\{ (500k)^{-1} \sum_{r=1}^{500} \sum_{j=1}^{k} (\widehat{Q}_{r,j} - Q_j)^2 \right\}^{1/2}.$$

Lower RMSE means more accurate point estimates. Average full width measures interval precision and is interpreted together with attained coverage.

4.1 PRECISION FROM COMPLEMENTARY DECODING

The matched three-stage comparison uses fused tree-OUE, balanced tree-OUE, and categorical randomized response (kRR) on the interval categories, with identical cohort sizes and privacy budgets. Fused three-cohort inference is the complete procedure in Algorithm 1. The two tree procedures isolate the effect of fusion; kRR replaces the reporting channel inside the same construction. Under $N(0,1)$, the public bracket is $[F^{-1}(0.02), F^{-1}(0.98)]$. Initial values interpolate linearly across that bracket, mapping the smallest and largest target probabilities to its endpoints. The defaults are the full Stage-1 average, $a = 2/3$, $\gamma_0 = 0.1(U - L)$, split $(0.35, 0.30, 0.35)$, and $\varpi_n = \{\log(n + e)\}^{-3/2}$. For channel m with Stage-2 envelope $B_{m,2k}$, the bandwidth is

$$h_{m,n} = \left\{ \frac{B_{m,2k}^2 \log\{2k(n+1)\}}{n_2} \right\}^{1/4},$$

using B_{2n} for tree-OUE and $(e^{\varepsilon} + 2k)/(e^{\varepsilon} - 1)$ for kRR. Appendix A.3 relates the normal-design gains to the sufficient gain condition and gives the remaining tuning details.

Table 1: End-to-end sample-size scaling at $\epsilon = 1$ under $N(0,1)$. Fused is the complete procedure in Algorithm 1; Balanced changes its tree decoder, and kRR changes its reporting channel. All methods use the same three-cohort allocation and public bracket. Each entry is simultaneous coverage followed by pooled RMSE; lower RMSE is better.

k	n	Fused	Balanced	kRR
31	160000	0.932; 0.0897	0.926; 0.1147	0.918; 0.1545
31	320000	0.940; 0.0602	0.944; 0.0766	0.928; 0.0949
63	320000	0.944; 0.0698	0.938; 0.0903	0.892; 0.2052
63	640000	0.948; 0.0490	0.936; 0.0625	0.902; 0.1340
127	640000	0.948; 0.0561	0.952; 0.0716	0.844; 0.3538
127	1280000	0.950; 0.0386	0.952; 0.0494	0.860; 0.2038

Table 1 separates grid growth from sample-size gains by doubling n at each fixed k. Fusion reduces RMSE by 21–23% relative to balanced decoding in all six settings, with coverage 93.2–95.0%. At $k = 127$ and $n = 1280000$, RMSE falls from 0.0494 to 0.0386 at essentially the same coverage (95.2% versus 95.0%). This precision gain uses the same reporting channel, cohort sizes, and privacy budget. At fixed k, doubling n reduces fused RMSE by 30–33%, consistent with the square-root inference-sample dependence of Corollary 3.3. The kRR procedure has substantially larger errors and undercovers at the larger grids.

Privacy and distribution comparisons in Appendix B retain the fusion gain: at $(n, k) = (320000, 63)$, fused RMSE decreases from 0.1490 to 0.0326 as ε increases from 0.5 to 2, with coverage 94.4–95.6%. Under the contaminated normal design at $(640000, 63, 1)$, fused RMSE is 0.0842, compared with 0.1053 for balanced decoding. Cohort allocation matters separately: increasing the inference share to

0.55 lowers error in regular settings but changes coverage from 0.944 to 0.896 at $\varepsilon = 0.5$ and from 0.938 to 0.872 under contamination. This allocation trades a smaller leading variance against larger learning errors.

4.2 Fused three-cohort inference and CDF-band inversion

A fixed spatial grid can use all n reports to estimate CDF values and then invert a joint CDF band. We implement this alternative with the same fused tree channel, anchor $u = 1/2$, and $d = k$ equally spaced inspection points on the common public range $[-3, 3]$. Both procedures use one $\varepsilon = 1$ report per user. Fused three-cohort inference instead learns the query locations and densities from its first two cohorts. Appendix A.4 specifies the band, inverse, and reporting-cost calculation; it uses the private CDF route of Liu et al. (2024) with our tree reporting channel and multiplier calibration.

Table 2: Complete procedures under $N(0, 1)$ on $[-3, 3]$, with one $\varepsilon = 1$ report per user. Each entry is (simultaneous coverage; RMSE; average full width). Both methods use all n users; the fused three-cohort method learns quantile locations and densities. Lower RMSE and width are preferable at comparable coverage.

n	k	Fused three-cohort	CDF inversion ($d = k$)
40000	31	0.942; 0.2704; 1.1112	1.000; 0.1064; 0.8216
160000	31	0.950; 0.1134; 0.5146	0.994; 0.0512; 0.5067
320000	63	0.918; 0.0809; 0.4326	0.992; 0.0395; 0.3556

Table 2 shows the resulting tradeoff. CDF inversion has smaller RMSE and narrower intervals in every setting, with conservative quantile coverage of 99.2–100%. At $(320000, 63)$, fused three-cohort coverage is 91.8%, with a 95% Wilson interval $[89.1, 93.9]\%$ across repetitions. Coverage remains 91.8% among runs without density clipping (Table 4). The underlying CDF band covers its spatial grid in 93.2–97.2% of repetitions; discrete inversion contributes the extra quantile conservatism.

The choice of spatial resolution also affects communication. Inverting exact CDF values on d equally spaced points has endpoint width $(U - L)/(d - 1)$. Refining from $d = 63$ to 127 at $(n, k) = (320000, 63)$ approximately halves this discretization width, but simulated widths remain 0.3556 and 0.3564. Mean report length rises from 24.0 to 39.3 bits, versus 28.6 for the three-cohort method. These upload counts include the privatized unary vector and level tag. The three-cohort method also updates public thresholds after each Stage-1 report; Appendix A.4 gives the separate query-delivery cost. The CDF release supports later probability queries; the three-cohort method learns locations for prespecified percentiles without choosing a spatial endpoint grid. The target list therefore determines its reporting dimensions, while the public range and endpoint resolution determine the CDF grid.

The comparison with corrected, fixed-block inference is complementary. With $J = 8$ restarted blocks, Student–Wishart and Student–Bonferroni attain 93.4–94.8% coverage; the fused three-cohort intervals are shorter, although their point errors are larger at $k = 31$ and 63. At $k = 127$, fused three-cohort and Student–Wishart widths are 0.3574 and 0.5902, a 39% reduction at similar coverage and RMSE. The block design and model-assisted correlation calibration are given in Appendix A.5; the fixed-J theory is in Appendix C.

5 California wage-percentile comparisons

We illustrate wage-percentile reporting with the 2024 one-year California person file from the American Community Survey Public Use Microdata Sample (ACS PUMS) (U.S. Census Bureau, 2025). The 90/50 and 50/10 ratios distinguish dispersion above and below the median (Autor et al., 2006). Each released record receives an additional local randomization.

We retain respondents aged 25–64 with positive wages, at least 40 weeks worked, and at least 30 usual weekly hours. We adjust annual wages to 2024 dollars and divide by annual hours, obtaining 123531 records. A random permutation precedes assignment to the three cohorts. We use log hourly wages, 19 probabilities from 0.05 to 0.95, $\varepsilon = 1$, split $(0.35, 0.30, 0.35)$, and public bracket $[\log 2, \log 500]$. The 294 records below this bracket and two above it remain in the outer interval categories.

For calibration, we treat unweighted log wages as independent draws from a continuous working distribution F. The target is $F^{-1}(\tau_j)$; the nominal 95% level refers to repeated independent streams from F, with fresh local randomization and multipliers, under Theorem 3.2's conditions. File quantiles serve as descriptive benchmarks. The released wages have ties, so this is a working-model application; Appendix A.6 gives the sampling, calibration, and local-density diagnostics. ACS population inference additionally requires survey weights and the PUMS replicate-weight design.

We apply the monotone interval refinement in Appendix A.2 on the log-wage scale and exponentiate its endpoints to obtain wage-quantile intervals. Write $[\widetilde{L}_j^I, \widetilde{U}_j^I]$ for the refined log-quantile interval. For $j < \ell$, the ratio of wage quantiles at τ_ℓ and τ_j inherits the interval $[\exp(\widetilde{L}_\ell^I - \widetilde{U}_j^I), \exp(\widetilde{U}_\ell^I - \widetilde{L}_j^I)]$. All such ratio intervals cover whenever the log-quantile band covers.

Table 3: Hourly-wage quantiles in 2024 dollars. Intervals use nominal 95% simultaneous calibration under the continuous working model.

τ	Empirical benchmark	Private estimate	95% simultaneous interval
0.10	14.40	13.44	[10.85, 16.66]
0.25	20.50	18.96	[16.27, 22.10]
0.50	34.17	30.71	[26.72, 35.30]
0.75	56.52	51.50	[43.69, 60.71]
0.90	90.30	100.37	[76.54, 131.61]

The private 90/50 and 50/10 ratios are 3.27 and 2.28, suggesting greater dispersion above the median. Transforming the simultaneous log-quantile band gives intervals $[2.17, 4.93]$ and $[1.60, 3.25]$, respectively; the 90/10 ratio is 7.47 with interval $[4.60, 12.13]$. This comparison uses the shared median on the log-wage scale. Writing $Q_\tau = F^{-1}(\tau)$ and $[L_\tau, U_\tau]$ for its simultaneous band endpoints, the contrast $D = Q_{0.90} + Q_{0.10} - 2Q_{0.50}$ inherits the interval

$$[L_{0.90} + L_{0.10} - 2U_{0.50},\ U_{0.90} + U_{0.10} - 2L_{0.50}].$$

It contains zero exactly when the rectangular-band intervals for the two log-ratios overlap; exponentiation preserves this overlap. Thus the reported band does not separate upper- and lower-tail dispersion. Figure 1 shows all 19 intervals: all five tabulated benchmarks lie inside, while the 40th-percentile benchmark (28.23) exceeds its upper endpoint (27.82).

6 Concluding remarks

This paper develops simultaneous quantile inference from one locally private report per user. The exact covariance of complementary scores yields a variance-minimizing fusion rule without additional privacy cost. Separating adaptive localization, density estimation, and fresh-score calibration gives a uniform one-step representation: first-order localization error is removed, and the remaining errors can be controlled on the scale of the maximum over the grid. This supports multiplier bands for polynomially growing grids under the stated conditions. Fusion improves matched three-stage estimates in the simulations, while comparison with CDF inversion clarifies the sample cost of learning query locations. The wage application shows how the joint band also supports percentile-ratio comparisons.

Two extensions merit further study. First, compact channels such as Hadamard response and local hashing (Acharya et al., 2019; Wang et al., 2017) suggest ways to reduce communication. Compression changes the dependence among decoded node scores, so the complementary-score covariance and optimal fusion weights would need to be derived anew. Uniform variance lower bounds and standardized moment bounds would then determine whether the localization rates and Gaussian approximation still permit comparably large grids. Second, continuously updating query locations and density estimates could make fuller use of an incoming stream. Time-uniform private quantile inference (Cai et al., 2025) offers a starting point, but a growing grid requires a one-step expansion uniform in both time and quantile. History-dependent scores would replace the present conditionally independent calibration sample, requiring control of accumulated conditional covariances and a Gaussian strong approximation or martingale confidence boundaries. Combining these bounds with uniform localization control would yield bands valid throughout online updating.

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

Appendices A.1 and A.2 give the complete algorithm and streaming identities. Appendix D.1 records the assumptions and Appendices H to J give the proofs. The numerical design and additional evidence are reported in Section 4 and Appendix B; Section 5 states the record-level application target and preprocessing. 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

Aamand, A., Boninsegna, F., Gentle, A., Imola, J., and Pagh, R. (2025). Lightweight protocols for distributed private quantile estimation. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 27–58.

Acharya, J., Sun, Z., and Zhang, H. (2019). Hadamard response: Estimating distributions privately, efficiently, and with little communication. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, PMLR 89, 1120–1129.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2006). The polarization of the U.S. labor market. *American Economic Review* **96**, 189–194.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics* **37**, 577–580.

Cai, L., Hu, Q., Sun, J., and Wu, S. (2025). Time-uniform and asymptotic confidence sequence of quantile under local differential privacy. *Advances in Neural Information Processing Systems* **38**, 114488–114520.

Cai, L., Hu, Q., and Wu, S. (2026). Federated learning of quantile inference under local differential privacy. *International Conference on Learning Representations*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics* **41**, 2786–2819.

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

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013b). Inference on counterfactual distributions. *Econometrica* **81**, 2205–2268.

Cormode, G., Kulkarni, T., and Srivastava, D. (2019). Answering range queries under local differential privacy. *Proceedings of the VLDB Endowment* **12**, 1126–1138.

Csörgő, M. (1983). *Quantile Processes with Statistical Applications*. CBMS-NSF Regional Conference Series in Applied Mathematics 42, SIAM, Philadelphia.

Dette, H. and Graw, C. (2024). Uncertainty quantification by block bootstrap for differentially private stochastic gradient descent. arXiv:2405.12553.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association* **113**, 182–201.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Kairouz, P., Oh, S., and Viswanath, P. (2014). Extremal mechanisms for local differential privacy. *Advances in Neural Information Processing Systems* **27**.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing* **40**, 793–826.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. Springer, New York.

Liu, Y. and Hu, Q. (2026). Locally private online quantile regression: Estimation and inference. arXiv:2607.05312.

Liu, Y., Hu, Q., Ding, L., and Kong, L. (2023). Online local differential private quantile inference via self-normalization. *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 21698–21714.

Liu, Y., Hu, Q., and Kong, L. (2024). Tuning-free estimation and inference of cumulative distribution function under local differential privacy. *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 31147–31164.

Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Annals of Probability* **18**, 1269–1283.

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

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

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

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

U.S. Census Bureau (2025). *2024 ACS 1-Year PUMS User Guide and Overview*. U.S. Department of Commerce, Washington, DC. Available at `https://www2.census.gov/programs-surveys/acs/tech_docs/pums/2024ACS_PUMS_User_Guide.pdf`.

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

Wang, T., Blocki, J., Li, N., and Jha, S. (2017). Locally differentially private protocols for frequency estimation. *26th USENIX Security Symposium*, 729–745.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**, 63–69.

Xie, J., Shi, E., Jiang, B., Kong, L., and He, X. (2025). Online differentially private inference in stochastic gradient descent. arXiv:2505.08227.

Zhu, W., Lou, Z., Wei, Z., and Wu, W. B. (2026). High confidence level inference is almost free using parallel stochastic optimization. arXiv:2401.09346, version 2.

The supplementary material first gives the algorithm, implementation, comparator protocols, and privacy/distribution experiments (Appendices A and B). Extra calibration theory follows in Appendix C, with full conditions in Appendices D and E. The remaining appendices supply the mechanism calculations, privacy argument, and inference proofs.

A Algorithm and implementation

A.1 The complete one-pass algorithm

Algorithm 1. Fused one-pass LDP inference for a growing quantile grid

Input: horizon n; target probabilities $\tau_1 < \cdots < \tau_{k_n}$; public bracket $[L, U]$; privacy level ε_n; split (ρ_1, ρ_2, ρ_3); gain (γ_0, a); bandwidth h_n; stabilization ϖ_n; level α; multiplier count $M \geq 1$. Process the cohorts consecutively, indexing users from one within each cohort.

1. Compute (n_1, n_2, n_3) from (2.6). Initialize deterministic $q_{0,j} \in [L, U]$ and $\overline{q}_{0,j} = 0$.
2. **For arrivals** $i = 1, \ldots, n_1$ **(Stage 1):** sort the current tagged thresholds, obtain one ε_n-LDP tree-OUE report, decode all fused scores with anchors τ_j, update (2.7), and update the full average by (A.1). Set $q_j^0 = \overline{q}_{n_1,j}$.
3. **For arrivals** $i = 1, \ldots, n_2$ **(Stage 2):** use the fixed tagged list $q_j^0 \pm h_n$, obtain one report per user, and update the $2k_n$ running score sums. Compute $\widetilde{f}_j$ and $\widehat{f}_j$ from (2.9)–(2.10).
4. Before Stage 3, independently generate multiplier streams $\{\xi_i^{(b)} : 1 \leq i \leq n_3, 1 \leq b \leq M\}$ from a public random seed.
5. **For arrivals** $i = 1, \ldots, n_3$ **(Stage 3):** use the fixed list q_j^0, obtain one report, and update for every j the running sums $\sum_i G_{i,j}^{(3)}$, $\sum_i (G_{i,j}^{(3)})^2$, and $\sum_i \xi_i^{(b)} G_{i,j}^{(3)}$ for $b \leq M$, together with $\sum_i \xi_i^{(b)}$.
6. Compute $\check{Q}_j$, $\widehat{\sigma}_{G,j}$, T_b^*, and the intervals in (2.11)–(2.14). Take the maximum over $\widehat{\mathcal{J}}_n = \{j : \widehat{\sigma}_{G,j} > 0\}$, using zero if this set is empty; return $[L, U]$ for excluded coordinates. Optionally apply the monotone interval refinement in Proposition K.1.

Stage assignment: the public arrival index alone determines an incoming user's stage.
Output: $\{\check{Q}_j, \mathcal{I}_{n,j} : 1 \leq j \leq k_n\}$.

The algorithm is online in the one-pass sense and uses memory independent of the stream length for fixed (M, k_n). Stage 1 adapts to past reports, whereas Stages 2 and 3 use frozen thresholds. The three boundaries and the coverage statement refer to a horizon fixed before data collection. Scalar confidence-sequence constructions use dynamically growing parallel paths for asymptotic time-uniform validity (Cai et al., 2025); the present fixed-terminal construction adds simultaneous calibration over a growing quantile grid.

A.2 Streaming memory and monotone post-processing

For each Stage-3 coordinate, a memory-bounded implementation maintains

$$S_{1,j} = \sum_{i=1}^{n_3} G_{i,j}^{(3)}, \qquad S_{2,j} = \sum_{i=1}^{n_3} (G_{i,j}^{(3)})^2.$$

Then

$$\overline{G}_j^{(3)} = S_{1,j}/n_3, \qquad \widehat{\sigma}_{G,j}^2 = S_{2,j}/n_3 - (S_{1,j}/n_3)^2.$$

For multiplier draw b, maintain

$$R_{b,j} = \sum_{i=1}^{n_3} \xi_i^{(b)} G_{i,j}^{(3)}, \qquad r_b = \sum_{i=1}^{n_3} \xi_i^{(b)}.$$

The centered numerator in (2.12) is

$$n_3^{-1/2}(R_{b,j} - r_b S_{1,j}/n_3).$$

These identities give $O(Mk_n)$ server memory, independent of the stream length. Vanilla tree-OUE uses $O(k_n)$ report bits. Compact alternatives such as optimized local hashing or Hadamard response require a channel-specific fusion covariance (Wang et al., 2017; Acharya et al., 2019). Floating-point implementations can compute variances with Welford's stable online recurrence. The simulation code returns $[L, U]$ for all coordinates if any estimated score variance is numerically zero or the covariance factorization fails.

Because $Q_1 \le \cdots \le Q_{k_n}$, define from raw intervals $[L_j^I, U_j^I]$

$$\widetilde{L}_j^I = \max_{\ell \le j} L_\ell^I, \qquad \widetilde{U}_j^I = \min_{\ell \ge j} U_\ell^I.$$

Proposition K.1 shows that this refinement preserves every simultaneous coverage event while enforcing the natural quantile ordering.

The full Stage-1 average is maintained without storing the path:

$$\overline{q}_{t,j} = \overline{q}_{t-1,j} + t^{-1}(q_{t,j} - \overline{q}_{t-1,j}). \tag{A.1}$$

The full average uses every projected iterate and preserves the entire Stage-1 sample.

For comparison, let $\overline{q}_{n_1,j}^{(b)} = (n_1 - b)^{-1} \sum_{t=b+1}^{n_1} q_{t,j}$. Proposition H.5 shows that its distance from the full average is at most $(U - L)b/n_1$. The two averages are therefore equivalent at rate r_{1n} when $b/n_1 = o(r_{1n})$.

The active set $\widehat{\mathcal{J}}_n$ and the zero-variance fallback are defined in Section 2.3. By Lemmas J.1 and J.2,

$$\mathbb{P}\{\widehat{\mathcal{J}}_n = \{1, \ldots, k_n\}\} \longrightarrow 1.$$

A.3 Default tuning

The split is $(0.35, 0.30, 0.35)$, Stage 1 uses its full average, and the gain is $0.1(U - L)t^{-2/3}$. Initial values linearly interpolate the target probabilities across the public bracket. The bandwidth is the mechanism-specific expression in Section 4 with multiplier one, and stabilization is $\varpi_n = \{\log(n + e)\}^{-3/2}$. At $M = 500$ and $\alpha = 0.05$, (2.13) uses rank 476 and the fixed-M coverage limit is $476/501 \approx 0.9501$.

For the normal design, write Φ and ϕ for the standard normal CDF and density. For both standard-normal designs, $L_F = \phi(0) = 1/\sqrt{2\pi}$ is admissible, so the main-design gain $\gamma_0 \approx 0.411$ and the CDF-comparison gain $\gamma_0 = 0.6$ satisfy the gain condition in Assumption D.4.

A.4 Fixed-grid CDF-band inversion

We compare the complete three-stage procedure with a fixed-grid CDF release under $N(0, 1)$, using the public bracket $[-3, 3]$, $\varepsilon = 1$, and $(n, k) = (40000, 31), (160000, 31), (320000, 63)$. Each method uses n users, one private report per user, 500 independent repetitions, and 500 multiplier draws per repetition. The direct procedure keeps the defaults in Appendix A.3. The CDF procedure places k equally spaced inspection points x_r across the bracket and uses all n users to estimate their CDF values. Its fused tree scores use the public anchor $u = 1/2$, which preserves unbiasedness without knowing $F(x_r)$. Thus the comparison includes the sample cost of localization and density estimation for the direct procedure.

Let $[\ell_r, u_r]$ be the studentized Gaussian-multiplier band for the CDF at these inspection points, clipped to $[0, 1]$. We invert it using

$$L_j^{\mathrm{CDF}} = \max\big(\{L\} \cup \{x_r : u_r < \tau_j\}\big), \qquad U_j^{\mathrm{CDF}} = \min\big(\{U\} \cup \{x_r : \ell_r \ge \tau_j\}\big).$$

On the simultaneous CDF coverage event, every resulting interval contains its target quantile. These endpoints retain the spatial discretization error; point estimates alone invert the piecewise-linear

interpolation of the isotonic CDF estimates. The target grid remains the k equally spaced probabilities on $[0.05, 0.95]$.

For d equally spaced inspection points including L and U, exact CDF values give a discrete inversion width of $(U - L)/(d - 1)$ for every target quantile strictly inside the bracket. Indeed, let r be the first grid index with $F(x_r) \geq \tau_j$. Then $F(x_{r-1}) < \tau_j$, so the endpoints above are x_{r-1} and x_r, including when $F(x_r) = \tau_j$. This calculation isolates spatial discretization from statistical uncertainty. The same simultaneous grid-CDF coverage event implies quantile coverage for every probability whose quantile is inside the bracket, so the released band also supports probabilities selected after collection.

For a report with d thresholds, the privatized unary vector has expected length $(2^{H_d+1} - 2)/H_d$ bits; transmitting the sampled level requires another $\lceil \log_2 H_d \rceil$ bits. The CDF procedure uses $d = k$ for every user; the direct procedure uses $d = k$ in Stages 1 and 3 and $d = 2k$ in Stage 2. Both fit the same maximum report-length allowance, with their actual mean lengths depending on the stage dimensions. Streaming multiplier storage is $O(Md)$ for a CDF grid with d inspection points and $O(Mk)$ for direct inference. For this simulation, both implementations compute the equivalent Gaussian critical value from the empirical covariance; the stored score matrices contain nk and n_3k doubles for CDF and direct inference, respectively.

The reporting counts include the privatized unary vector and level tag, excluding public-query broadcasts and transport framing. At $k = 31$, direct and $d = k$ CDF inversion use 18.0 and 15.4 bits on average; at $k = 63$, they use 28.6 and 24.0. For a refined CDF grid with $d = 127$, the mean is 39.3 bits. The 0.70/0.30 weighted mean of the k- and $2k$-threshold report lengths gives the direct cost under the default allocation. The number of users and privacy budget are fixed across these comparisons; average communication follows the reporting dimensions.

For $(n, k) = (40000, 31), (160000, 31), (320000, 63)$, direct density clipping occurs in 26.0%, 4.8%, and 2.0% of repetitions; neither procedure produces fallback or empty intervals. The covariance-based simulation implementations store nk and n_3k scores for CDF and direct inference. The reported timings measure these covariance-based implementations; Appendix A.2 gives the $O(Mk)$ streaming construction.

Algorithm 1 serves users sequentially: the server publishes the current ordered threshold list before each Stage-1 report and updates it after that report. The Stage-2 and Stage-3 lists are fixed within their cohorts. Sending each user the full list therefore requires $k(n + n_2)$ threshold values in total. With 64-bit values, this is about 210 MB at $(n, k) = (320000, 63)$, in addition to the private uploads. Counting each public list publication once gives $k(n_1 + 3)$ values; delivery to individual users still depends on the broadcast or retrieval protocol. An equally spaced CDF grid is specified by its endpoints and size and can be reconstructed by each user. The reported simulation times use in-memory queries and measure no network transport. Thus upload length, query distribution, and computation are separate deployment costs.

Table 4 quantifies uncertainty in the direct procedure's coverage and separates repetitions with and without density clipping. The Wilson intervals describe Monte Carlo uncertainty in the coverage probability. At $(320000, 63)$, 40 of the 490 unclipped repetitions miss at least one target, giving essentially the same 91.8% coverage as the full set. Clipping alone therefore does not explain this deficit. The stratification is descriptive: clipping is an outcome of the same learning stages whose errors enter the one-step remainder.

Table 4: Fused three-cohort inference on $[-3, 3]$ under $N(0, 1)$ at $\varepsilon = 1$. All-run entries give coverage and its 95% Wilson interval across 500 repetitions. The last two columns give covered repetitions divided by repetitions in that density-clipping group.

n	k	All runs	No clipping	Clipping
40000	31	0.942 [0.918, 0.959]	349/370	122/130
160000	31	0.950 [0.927, 0.966]	452/476	23/24
320000	63	0.918 [0.891, 0.939]	450/490	9/10

Table 5: Corrected block procedures versus fused three-cohort inference (Algorithm 1) at $\epsilon = 1$ under $N(0, 1)$. Restarted methods use $J = 8$ and balanced tree-OUE; the complete designs differ as specified in the text. Width is the mean full interval width.

n	k	Method	Coverage	RMSE	Width
160000	31	Fused three-cohort	0.932	0.0897	0.4777
		Student–Wishart	0.938	0.0764	0.6367
		Student–Bonferroni	0.948	0.0764	0.6432
320000	63	Fused three-cohort	0.944	0.0698	0.4178
		Student–Wishart	0.934	0.0662	0.6165
		Student–Bonferroni	0.946	0.0662	0.6240
640000	127	Fused three-cohort	0.948	0.0561	0.3574
		Student–Wishart	0.944	0.0570	0.5902
		Student–Bonferroni	0.944	0.0570	0.5971

A.5 Corrected restarted-block comparisons

The block comparisons use $N(0, 1)$, $\varepsilon = 1$, and $(n, k) = (160000, 31), (320000, 63), (640000, 127)$, with 500 repetitions and 500 calibration draws where needed. They divide all users into $J = 8$ equal blocks and use balanced tree-OUE. Each block initializes by linear interpolation across $[\Phi^{-1}(0.01), \Phi^{-1}(0.99)]$, restarts the gain $0.1(U - L)t^{-2/3}$, and averages its second half. Student–Wishart and Student–Bonferroni share the resulting point estimator; fused three-cohort inference instead uses the main three-cohort design and its $[\Phi^{-1}(0.02), \Phi^{-1}(0.98)]$ bracket. The comparison includes these decoder, bracket, path, and allocation choices.

Student–Wishart uses the public, model-based balanced tree score correlation in the Gaussian block approximation. It is a model-assisted comparator; Student–Bonferroni uses the corresponding marginal t tails. Appendix C defines both references. Table 5 retains the coverage–error–width tradeoff: fused three-cohort bands are shorter in all settings, while restarted estimates have smaller RMSE at $k = 31$ and 63 and similar RMSE at $k = 127$.

A.6 Application model and diagnostics

The application uses a continuous i.i.d. working model for the unweighted log wages. Its inferential target is the model quantile vector, and its repeated experiment draws a new stream of eligible records before applying the three-stage procedure and fresh private randomization. The asymptotic coverage statement applies when the model meets Assumptions D.1 to D.6. For the fixed released file, the empirical quantiles are descriptive comparison points; their inclusion in one realized band is a different event from coverage of the model quantiles across independent streams.

The records are randomly permuted before local randomization and then assigned by arrival index to cohorts of sizes 43235, 37059, and 43237. The gain constant is $0.1 \log(500/2) \approx 0.552$.

The 123531 log wages take 7260 distinct values. The largest empirical point mass is 3.2%; masses at the 19 empirical target quantiles range from less than 0.01% to 2.7%. On the log scale, windows of radius $h_n \approx 0.638$ around those quantiles contain 1751–3385 distinct values. The 19 estimated densities range from 0.123 to 0.507, with no density clipping or variance fallback. The finite difference therefore averages over many recorded values, while smooth local density remains a working approximation.

Of the 19 file quantiles, 18 lie in the realized band. The exception is the 40th-percentile benchmark 28.23, whose interval is $[22.41, 27.82]$. All five benchmarks in Table 3 lie inside. The ratio-interval transformation preserves each joint inclusion event algebraically; its nominal model interpretation inherits the same sampling and smoothness assumptions as the original band.

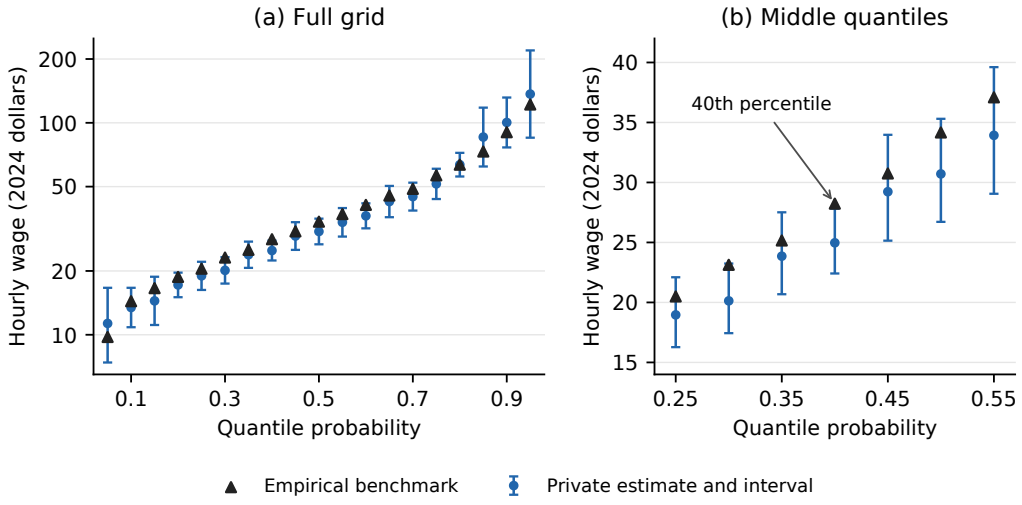


Figure 1: Private hourly-wage quantiles and nominal 95% simultaneous intervals for the 19-point ACS grid in Section 5. The left panel uses a logarithmic wage axis; the right enlarges the middle quantiles on a linear axis. Black triangles are empirical file benchmarks. At the 40th percentile, the benchmark lies above its interval. Calibration targets the continuous working-model quantiles; the plotted benchmarks describe the fixed released file.

B Privacy and distribution comparisons

These experiments retain the three-stage design, target probabilities, and mechanism-specific tuning in Section 4. Each cell uses 500 repetitions and 500 multipliers. The public bracket is $[F^{-1}(0.02), F^{-1}(0.98)]$ for the relevant distribution, using population-level range information. Coverage is for the complete target grid, and each table pairs it with pooled RMSE.

At fixed $(n, k) = (320000, 63)$ under $N(0, 1)$, varying ε isolates the privacy effect within each procedure. Table 6 shows a clear precision gain from weaker privacy: fused RMSE falls from 0.1490 to 0.0326, with coverage 94.4–95.6%. Fused RMSE is smaller than balanced RMSE at every privacy level.

Table 6: End-to-end privacy comparison at $n = 320000$ and $k = 63$ under $N(0, 1)$. Each method entry is simultaneous coverage followed by pooled RMSE.

ϵ	Fused	Balanced	kRR
0.5	0.944; 0.1490	0.932; 0.2193	0.874; 1.1007
1.0	0.944; 0.0698	0.938; 0.0903	0.892; 0.2052
2.0	0.956; 0.0326	0.932; 0.0402	0.934; 0.0474

At $(n, k, \varepsilon) = (640000, 63, 1)$, we compare $N(0, 1)$, a variance-standardized t_3 distribution, and $0.9N(0, 1) + 0.1N(0, 4^2)$. Fused coverage is 93.8–94.8% across the three distributions (Table 7). Contamination increases its RMSE to 0.0842, consistent with the importance of local density for quantile precision; fusion still improves on balanced decoding, whose RMSE is 0.1053. kRR has appreciably larger error and lower coverage under both heavy tails and contamination.

Table 7: End-to-end distribution comparison at $n = 640000$, $k = 63$, and $\epsilon = 1$. Each method entry is simultaneous coverage followed by pooled RMSE.

Distribution	Fused	Balanced	kRR
Normal	0.948; 0.0490	0.936; 0.0625	0.902; 0.1340
t_3	0.948; 0.0483	0.932; 0.0597	0.862; 0.1182
Contamination	0.938; 0.0842	0.932; 0.1053	0.844; 0.2377

C Fixed-Chain Calibration Results

C.1 Restarted-chain procedures

This subsection defines the blockwise procedures used in the calibration comparison. Let J be the number of blocks and $\widehat{Q}_b = (\widehat{Q}_{b,1}, \ldots, \widehat{Q}_{b,k_n})^\top$ the averaged private stochastic-approximation output from block b. The block aggregate and coordinate scales are

$$\widetilde{Q} = J^{-1} \sum_{b=1}^{J} \widehat{Q}_b, \qquad S_j^2 = J^{-1} \sum_{b=1}^{J} (\widehat{Q}_{b,j} - \widetilde{Q}_j)^2.$$

The raw chain multiplier statistic is

$$T^*_{\mathrm{chain}} = \max_{j \le k_n} \left| \frac{J^{-1/2} \sum_{b=1}^{J} \xi_b (\widehat{Q}_{b,j} - \widetilde{Q}_j)}{S_j} \right|.$$

This raw multiplier approximates the large-J regime in which block bias is small relative to $S_j/\sqrt{J}$.

Under a multivariate normal block model, the sample mean is independent of the Wishart sample covariance. The fixed-J Student–Wishart comparator simulates this pivot using the public balanced tree-OUE score correlation, thereby retaining the random coordinate scales and cross-coordinate dependence. For restarted stochastic-approximation outputs, it gives the corresponding first-order Gaussian calibration.

More explicitly, let $k = k_n$, let R be the public score correlation matrix, and put $\nu = J - 1$. Let $V \sim \mathrm{Wishart}_k(R, \nu)$ be independent of $Z \sim N_k(0, R)$. Orthogonally separating the block mean from the ν residual contrasts gives the exact Gaussian-model maximum

$$\max_{j \le k} \frac{|Z_j|}{\sqrt{V_{jj}/\nu}}.$$

This representation remains valid for singular R. Its critical value depends on R: if $R = \mathbf{1}\mathbf{1}^\top$, all coordinates coincide and the maximum has the $|t_\nu|$ distribution for every k. The polynomial rate in Proposition C.2 therefore applies to the Student–Bonferroni calibration; the Student–Wishart rate is determined by the correlation matrix R.

The marginal Student–Bonferroni safeguard uses

$$\widetilde{Q}_j \pm t_{J-1,1-\alpha/(2k_n)} \frac{s_j}{\sqrt{J}}, \qquad s_j^2 = (J-1)^{-1} \sum_{b=1}^{J} (\widehat{Q}_{b,j} - \widetilde{Q}_j)^2,$$

Together with the raw multiplier, these intervals distinguish fixed-J tail calibration from blockwise estimation bias. Their polynomial critical-value growth is given in Proposition C.2.

The common block bias and fixed-J target distribution remain unchanged as the number of multiplier draws increases. Increasing the number of chains with the horizon yields the distinct time-uniform construction of Cai et al. (2025).

C.2 Calibration with a fixed number of restarted chains

Proposition C.1 (Fixed-J Gaussian-multiplier mismatch)**.** *Let $W_1, \ldots, W_J$ be i.i.d. $N(\theta, \sigma^2)$ with fixed $J \ge 2$, and let*

$$\overline{W} = J^{-1} \sum_{b=1}^{J} W_b, \qquad S_J^2 = J^{-1} \sum_{b=1}^{J} (W_b - \overline{W})^2.$$

Then

$$\frac{\sqrt{J-1}(\overline{W}-\theta)}{S_J} \sim t_{J-1}.$$

In contrast, conditional on the observations,

$$\frac{J^{-1/2}\sum_{b=1}^{J}\xi_b(W_b-\overline{W})}{S_J} \sim N(0,1).$$

The discrepancy persists as the within-block sample size grows whenever J remains fixed.

Proposition C.2 (Polynomial Student–Bonferroni growth)**.** *For fixed degrees of freedom ν and fixed $\alpha \in (0,1)$,*

$$t_{\nu,1-\alpha/(2k)} \asymp k^{1/\nu} \qquad (k \to \infty). \tag{C.1}$$

Thus a fixed-J Student–Bonferroni band has a critical value of polynomial order $k^{1/(J-1)}$, and Holm's (1979) most stringent step has the same fixed-degree tail. A Gaussian maximum over k standardized coordinates has critical value at most of order $\sqrt{\log k}$ under arbitrary correlation.

The two propositions separate calibration from estimation. Student–Bonferroni has the polynomial critical-value growth in (C.1). Student–Wishart instead uses the public first-order score correlation induced by balanced tree OUE, so its critical value depends on that correlation matrix. Both corrected fixed-J procedures produce wider intervals in the numerical comparisons, whereas the user-level multiplier uses n_3 independent score vectors.

C.3 Connections with stochastic-approximation inference

The construction also relates to uncertainty quantification for stochastic approximation. HiGrad creates hierarchically split trajectories and t-based contrasts (Su and Zhu, 2023), while Zhu et al. (2026) analyze high-confidence t inference from a small number of independent stochastic-optimization runs. Dette and Graw (2024) develop block-bootstrap inference for stochastic gradients under pure local differential privacy. Xie et al. (2025) use locally perturbed gradients with Gaussian differential privacy and derive a functional central limit theorem. Growing-grid inference additionally requires an approximation for the maximum over quantile coordinates. We obtain it by using the first path for localization and an independent cohort of user-level scores for inference. The private channel builds on randomized response, frequency estimation, and range-query mechanisms (Warner, 1965; Kasiviswanathan et al., 2011; Kairouz et al., 2014; Wang et al., 2017; Duchi et al., 2018; Acharya et al., 2019; Cormode et al., 2019); the localization proof uses classical stochastic approximation and averaging (Robbins and Monro, 1951; Ruppert, 1988; Polyak and Juditsky, 1992; Kushner and Yin, 2003); and the final calibration invokes high-dimensional Gaussian and multiplier approximations (Chernozhukov et al., 2013a; 2017).

D Complete Regularity and Rate Conditions

D.1 Assumptions and rate quantities

Write $\ell_n = \log\{2k_n(n+1)\}$. For a threshold list of length d, use H_d, ν_{ε_n}, and $\mathcal{B}_{d,\varepsilon_n}$ from (2.1). Set

$$B_{1n} = \mathcal{B}_{k_n,\varepsilon_n}, \qquad B_{2n} = \mathcal{B}_{2k_n,\varepsilon_n}.$$

Assumption D.1 (Sampling, targets, and interiority)**.** *The observations are i.i.d. with continuous CDF F. There are constants $\delta_0 > 0$ and $0 < \tau_- < \tau_+ < 1$ such that, for every n,*

$$\tau_- \le \tau_{1,n} < \cdots < \tau_{k_n,n} \le \tau_+, \qquad Q_{j,n} \in [L+3\delta_0, U-3\delta_0] \quad (1 \le j \le k_n).$$

Every user participates exactly once and uses independent local randomization conditional on the public transcript.

Assumption D.2 (Local quantile regularity)**.** *For some $\beta \in (0,1]$, the CDF is absolutely continuous on*

$$\mathcal{N}_n = \bigcup_{j \le k_n} [Q_{j,n} - 2\delta_0, Q_{j,n} + 2\delta_0]$$

with density f satisfying, uniformly on $\mathcal{N}_n$,

$$0 < f_- \le f(x) \le f_+ < \infty, \qquad |f(x) - f(y)| \le L_f|x-y|^\beta.$$

The constants do not depend on n.

Positivity and local continuity of the density are standard regular-quantile conditions for Bahadur and one-step expansions (Bahadur, 1966; van der Vaart, 1998); closely related local conditions are used for online locally private quantile inference in Liu et al. (2023). The Hölder bound gives a uniform remainder over the growing target grid.

To control Stage 1 from any initialization in the projection bracket, we use the following secant condition.

Assumption D.3 (Uniform Stage-1 identification)**.** *There are constants $0 < \mu \le L_F < \infty$ such that for every n, every $j \le k_n$, and every $q \in [L, U]$,*

$$\mu|q - Q_{j,n}| \le |F(q) - \tau_{j,n}| \le L_F|q - Q_{j,n}|, \qquad (q - Q_{j,n})\{F(q) - \tau_{j,n}\} \ge 0. \tag{D.1}$$

Assumption D.3 is the monotone root-identification condition for the projected Robbins–Monro step on $[L, U]$. Together with the decreasing gain in Assumption D.4, it follows the standard stochastic-approximation setup of Robbins and Monro (1951); Polyak and Juditsky (1992); Kushner and Yin (2003). Appendix E.1 shows how local density regularity implies this condition on the public bracket.

The index b_n below marks entry into the local neighborhood in the analysis; the implemented full average retains every Stage-1 iterate.

Assumption D.4 (Gain, stage sizes, and localization)**.** *The gain is $\gamma_t = \gamma_0 t^{-a}$ with $a \in (1/2, 1)$ and fixed*

$$0 < \gamma_0 \le \frac{1}{2L_F}.$$

The initial values are deterministic in $[L, U]$. There is a deterministic integer $b_n \in \{1, \dots, n_1 - 1\}$ satisfying

$$b_n/n_1 \to 0, \tag{D.2}$$

$$\exp\left(-\frac{\mu}{2}\sum_{t=1}^{b_n}\gamma_t\right)(U - L) = o\{\min(r_{1n}, \delta_0)\}, \tag{D.3}$$

$$B_{1n}\left\{\sqrt{\gamma_{b_n}\ell_n} + \gamma_{b_n}\ell_n\right\} = o(\delta_0), \tag{D.4}$$

Here r_{1n} is defined in (D.5) *below. Together with Assumptions D.1 and D.3, conditions* (D.2)–(D.4) *give uniform entry into the local neighborhood and eventual inactivity of the projection; a direct choice of b_n is given in Lemma H.1.*

The explicit preliminary rate is

$$\begin{aligned} r_{1n} = \frac{b_n}{n_1} + B_{1n}\sqrt{\frac{\ell_n}{n_1}} + B_{1n}n_1^{a/2-1}\sqrt{\ell_n} \\ + B_{1n}^{1+\beta}n_1^{-a(1+\beta)/2}\ell_n^{(1+\beta)/2}. \end{aligned} \tag{D.5}$$

The four terms correspond to the initial segment, the averaged martingale, the Abel endpoint, and the nonlinear Hölder remainder.

Assumption D.5 (Density stage)**.** *The bandwidth satisfies $h_n \to 0$, $h_n < \delta_0$, and*

$$d_n = r_{1n}^{\beta} + h_n^{\beta} + \frac{B_{2n}}{h_n}\left(\sqrt{\frac{\ell_n}{n_2}} + \frac{\ell_n}{n_2}\right) \longrightarrow 0.$$

The deterministic clipping sequence obeys $\varpi_n \downarrow 0$ and $\varpi_n^{-1} \to \infty$.

Let $p_j^0 = F(q_j^0)$ and, for a Stage-3 observation, define

$$Y_{i,j} = G_{i,j}^{(3)} - p_j^0, \qquad \sigma_{j,n}^2 = \mathbb{E}(Y_{i,j}^2 \mid \mathcal{F}_{n_1+n_2}).$$

Let

$$\underline{\sigma}_n = \min_{j \le k_n}\sigma_{j,n}, \qquad K_n = \frac{2B_{1n}}{\underline{\sigma}_n}.$$

By Proposition 2.1, $|Y_{i,j}|/\sigma_{j,n} \le K_n$ almost surely conditional on the first two stages. The resulting support and variance bounds are established in Lemma J.1.

Assumption D.6 (High-dimensional approximation and one-step remainder)**.** *The following random rates converge to zero in probability:*

$$\left(\frac{K_n^2 \ell_n^7}{n_3}\right)^{1/6} = o_{\mathbb{P}}(1), \tag{D.6}$$

$$\frac{\sqrt{n_3 \ell_n}}{\underline{\sigma}_n}\{r_{1n} d_n + r_{1n}^{1+\beta}\} = o_{\mathbb{P}}(1). \tag{D.7}$$

For an empirical multiplier quantile based on M_n draws, assume $M_n \to \infty$. The exact conditional multiplier quantile is the theoretical reference.

The first condition is the Gaussian approximation condition for bounded high-dimensional vectors in Chernozhukov et al. (2017); the second makes the one-step remainder negligible on the maximum scale. Corollary 3.3 gives a simple growth regime that implies both conditions.

E Interpretation of the assumptions

This appendix gives primitive sufficient conditions and useful alternatives for the main assumptions.

E.1 Density regularity and identification

The lower bound $f_- > 0$ on a neighborhood of the target range identifies each quantile locally and converts CDF error into quantile error. The β-Hölder condition controls the density bias in (I.2) and the one-step remainder in (I.5). A locally Lipschitz density gives $\beta = 1$; a common modulus of continuity can be used in the same arguments when explicit polynomial rates are less important.

In applications, the minimum estimated density, the number of active clipping operations, and bandwidth sensitivity help identify quantiles with weak local information. Extreme targets with vanishing density require the corresponding nonregular quantile normalization.

Assumption D.3 is a secant condition on the public bracket $[L, U]$. A primitive sufficient condition is absolute continuity on $[L, U]$ with

$$0 < \mu \le f(x) \le L_F < \infty \quad \text{for almost every } x \in [L, U].$$

Indeed,

$$F(q) - F(Q_j) = \int_{Q_j}^{q} f(x)\, dx$$

gives (D.1).

The local regularity in Assumption D.2 also yields the global secant bounds on the fixed bracket. For $|q - Q_j| \le \delta_0$, integration gives constants f_- and f_+. If $q \ge Q_j + \delta_0$, monotonicity gives

$$\frac{F(q) - F(Q_j)}{q - Q_j} \ge \frac{f_- \delta_0}{U - L}, \qquad \frac{F(q) - F(Q_j)}{q - Q_j} \le \frac{1}{\delta_0},$$

and the same bounds hold on the left of Q_j. Thus (D.1) holds with $\mu = \min\{f_-, f_- \delta_0/(U - L)\}$ and $L_F = \max\{f_+, 1/\delta_0\}$. The positive burn-in conditions in Assumption D.4 then control early private-noise excursions.

Public scientific range information or an independently privatized range-estimation cohort can supply $[L, U]$. The Stage-1 projection frequency and the distance of final iterates from the bracket endpoints summarize its adequacy.

E.2 Target spacing, cohort sizes, and privacy

The target spacing may decrease with n. Stable tags preserve target identities when close thresholds cross during Stage 1 or coincide after finite-precision rounding, and the score remains unbiased at each tagged threshold. The empirical multiplier incorporates the resulting cross-quantile correlation.

The main statement uses fixed positive fractions. More generally, the proofs require

$$n_1 \to \infty, \qquad n_2 \to \infty, \qquad n_3 \to \infty,$$

together with Assumptions D.4 to D.6. In particular, $n_2/n \to 0$ is allowed if

$$\frac{B_{2n}}{h_n}\sqrt{\frac{\ell_n}{n_2}} \to 0$$

and the one-step remainder condition hold. Increasing n_3 improves first-order precision after Stage 1 has localized the targets and Stage 2 has estimated their densities accurately. This is the tradeoff illustrated by the allocation results summarized in Section 4.1: smaller learning cohorts can increase the remainder even while n_3 reduces the leading variance.

The main theorems allow ε_n to vary through ν_{ε_n} and the score envelope. For small ε_n,

$$\nu_{\varepsilon_n} \asymp \varepsilon_n^{-2}, \qquad \mathcal{B}_{d,\varepsilon_n} \asymp H_d \varepsilon_n^{-1}.$$

These quantities display the effective private sample size in each rate. Corollary 3.3 specializes them to fixed ε for a simple grid-growth condition.

E.3 Proof ingredients

The privacy calculation and fused-score moments are specific to the proposed mechanism. The Abel decomposition and one-step algebra are established directly. Lemma H.1 specializes a classical stopped Lyapunov argument for projected stochastic approximation (Robbins and Monro, 1951; Ruppert, 1988; Polyak and Juditsky, 1992; Kushner and Yin, 2003). The Gaussian approximation, multiplier bootstrap, and anti-concentration inputs follow Chernozhukov et al. (2013a; 2017), with their conditional moment and dimension requirements verified in Lemma J.3. The Gaussian mean–sample covariance independence in Proposition C.1 is standard multivariate analysis (Anderson, 2003).

F Tree-OUE Moments and Fusion

This appendix proves Proposition 2.1 from the main text. Throughout, the public threshold list, its tree, and the anchor are fixed.

F.1 Node moments

Let $I_v = \mathbf{1}\{C(X) \text{ is a descendant leaf of } v\}$. Condition on $L^\star = \ell(v)$ for the following bit moments; the input bit is I_v. Put

$$\eta_v = \frac{Z_v - q_\varepsilon}{p_\varepsilon - q_\varepsilon}.$$

If $I_v = 0$, then $Z_v \sim \mathrm{Bernoulli}(q_\varepsilon)$ and

$$\mathbb{E}(\eta_v \mid I_v = 0) = 0,$$
$$\mathbb{E}(\eta_v^2 \mid I_v = 0) = \frac{q_\varepsilon(1-q_\varepsilon)}{(p_\varepsilon - q_\varepsilon)^2} = \frac{4e^\varepsilon}{(e^\varepsilon - 1)^2} = \nu_\varepsilon.$$

If $I_v = 1$, then $Z_v \sim \mathrm{Bernoulli}(p_\varepsilon)$ and

$$\mathbb{E}(\eta_v \mid I_v = 1) = 1,$$
$$\mathbb{E}(\eta_v^2 \mid I_v = 1) = 1 + \frac{p_\varepsilon(1-p_\varepsilon)}{(p_\varepsilon - q_\varepsilon)^2} = \nu_\varepsilon + 2.$$

The last identity follows by direct substitution of $p_\varepsilon = 1/2$ and $q_\varepsilon = (e^\varepsilon + 1)^{-1}$.

Averaging the sampled-level moments over the uniform $L^\star$ gives

$$\mathbb{E}(\psi_v \mid X) = H_d \mathbb{P}\{L^\star = \ell(v)\}\mathbb{E}(\eta_v \mid X, L^\star = \ell(v)) = I_v, \tag{F.1}$$
$$\mathbb{E}(\psi_v^2 \mid X) = H_d^2 \mathbb{P}\{L^\star = \ell(v)\}\mathbb{E}(\eta_v^2 \mid X, L^\star = \ell(v)) = H_d(\nu_\varepsilon + 2I_v). \tag{F.2}$$

If $v \neq w$ are disjoint tree nodes, then

$$\mathbb{E}(\psi_v \psi_w \mid X) = 0. \tag{F.3}$$

Indeed, if their levels differ, the two level indicators cannot both equal one. If their levels agree, their unary bits are conditionally independent and at most one input bit equals one. The centered decoder for every zero input has conditional mean zero, so the product has conditional mean zero.

Finally, since $|Z_v - q_\varepsilon| \leq 1 - q_\varepsilon$ and

$$\frac{1 - q_\varepsilon}{p_\varepsilon - q_\varepsilon} = \frac{2e^\varepsilon}{e^\varepsilon - 1},$$

we have

$$|\psi_v| \leq \frac{2H_d e^\varepsilon}{e^\varepsilon - 1}.$$

F.2 Prefix, suffix, and fusion

Let $p_r = F(x_{(r)})$. Exactly one node in $\mathcal{D}_r^-$ contains the realized leaf when $X \leq x_{(r)}$, and none does otherwise. Summing (F.1) gives

$$\mathbb{E}(A_r \mid X) = \mathbf{1}\{X \leq x_{(r)}\}, \qquad \mathbb{E}A_r = p_r.$$

Using (F.2) and (F.3),

$$\begin{aligned}\mathbb{E}A_r^2 &= \sum_{v \in \mathcal{D}_r^-} \mathbb{E}\psi_v^2 = H_d w_r^- \nu_\varepsilon + 2H_d \sum_{v \in \mathcal{D}_r^-} \mathbb{P}(I_v = 1) \\ &= H_d w_r^- \nu_\varepsilon + 2H_d p_r.\end{aligned}$$

Therefore

$$\mathrm{Var}(A_r) = H_d w_r^- \nu_\varepsilon + 2H_d p_r - p_r^2.$$

Write $S_r = \sum_{v \in \mathcal{D}_r^+} \psi_v$, so $B_r = 1 - S_r$. The same argument gives

$$\mathbb{E}S_r = 1 - p_r, \qquad \mathbb{E}S_r^2 = H_d w_r^+ \nu_\varepsilon + 2H_d(1 - p_r),$$

and hence

$$\mathrm{Var}(B_r) = \mathrm{Var}(S_r) = H_d w_r^+ \nu_\varepsilon + 2H_d(1 - p_r) - (1 - p_r)^2.$$

Every prefix node is disjoint from every suffix node, so (F.3) yields $\mathbb{E}(A_r S_r) = 0$. It follows that

$$\mathrm{Cov}(A_r, B_r) = \mathrm{Cov}(A_r, 1 - S_r) = -\{\mathbb{E}(A_r S_r) - \mathbb{E}A_r \mathbb{E}S_r\} = p_r(1 - p_r).$$

For an arbitrary scalar λ, the estimator $\lambda A_r + (1 - \lambda)B_r$ is unbiased, with variance

$$\mathrm{Var}\{\lambda A_r + (1 - \lambda)B_r\} = \lambda^2 \mathrm{Var}(A_r) + (1 - \lambda)^2 \mathrm{Var}(B_r) + 2\lambda(1 - \lambda)p_r(1 - p_r), \tag{F.4}$$

Set $C_r = p_r(1 - p_r)$ and

$$a = a_r^-(p_r) = \mathrm{Var}(A_r) - C_r, \qquad b = a_r^+(p_r) = \mathrm{Var}(B_r) - C_r.$$

Completing the square gives

$$\mathrm{Var}\{\lambda A_r + (1 - \lambda)B_r\} = C_r + a\lambda^2 + b(1 - \lambda)^2 = C_r + \frac{ab}{a + b} + (a + b)\left(\lambda - \frac{b}{a + b}\right)^2. \tag{F.5}$$

Both a and b are strictly positive. The minimizer is $b/(a + b) = \lambda_r^\circ(p_r)$ and the minimum is (2.3). Since minimizing over $[0, 1]$ cannot be worse than evaluating at zero or one, the variance dominance follows.

The sum $D_r = a_r^-(u) + a_r^+(u) = H_d(w_r^- + w_r^+)\nu_\varepsilon + 2H_d - 1$ does not depend on u, and

$$\lambda_r^\circ(u) - \lambda_r^\circ(p_r) = -\frac{(2H_d - 1)(u - p_r)}{D_r}.$$

Substitution into (F.5) gives (2.4). Since $D_r \geq 2H_d - 1$, it also gives (2.5). Finally, at the sampled level each canonical decomposition contributes at most one node score. Hence $|A_r| \leq 2H_d e^\varepsilon/(e^\varepsilon - 1)$ and $|B_r| \leq 1 + 2H_d e^\varepsilon/(e^\varepsilon - 1)$. Convexity of the fused score proves the claimed envelope. □

G PRIVACY PROOF

Sequential privacy guarantee. Fix a public threshold list and a sampled level ℓ. Two possible interval categories produce unary vectors that are either identical or differ in exactly two coordinates: one coordinate changes from one to zero and another from zero to one. For one changed pair and any released bit pair (z_1, z_2), the largest likelihood ratio is

$$\frac{p_\varepsilon(1-q_\varepsilon)}{q_\varepsilon(1-p_\varepsilon)} = \frac{1-q_\varepsilon}{q_\varepsilon} = e^\varepsilon.$$

All unchanged bit factors cancel. Thus the conditional OUE release at level ℓ is ε-LDP. The level is sampled independently of the private category, so mixing over levels preserves the same bound.

Now consider the sequential transcript $Z_{1:n}$. Let two data streams differ only at user i. For a fixed realized transcript, the chain rule factorizes its likelihood into conditional kernels. The factors before i coincide. At i, the public query is the same function of the common preceding transcript and the conditional likelihood ratio is at most e^{ε_n}. Conditional on the realized report Z_i and the subsequent common transcript prefix, all later public queries are again the same deterministic functions of that prefix. Because later users' raw values agree, every later conditional kernel cancels in the likelihood ratio. Hence the complete transcript ratio is at most e^{ε_n}.

All server computations after receipt of the reports—sorting, score decoding, projection, averaging, density estimation, multiplier generation, and interval refinement—are measurable post-processing and cannot increase the privacy loss. Finally, a user belongs to exactly one cohort, so there is no across-stage composition for that user. This proves the privacy guarantee. □

H STAGE-1 LOCALIZATION AND FULL-AVERAGE EXPANSION

We first record the filtration and a localization lemma. For Stage 1, let $\mathcal{F}_t$ contain all observations only through their released reports up to time t, together with all public randomization. Put

$$e_{t,j} = q_{t,j} - Q_{j,n}, \qquad \zeta_{t,j} = G^{(1)}_{t,j} - F(q_{t-1,j}).$$

Then

$$\mathbb{E}(\zeta_{t,j} \mid \mathcal{F}_{t-1}) = 0, \qquad |\zeta_{t,j}| \le 2B_{1n}. \tag{H.1}$$

The bound uses $0 \le F \le 1$ and $|G^{(1)}_{t,j}| \le B_{1n}$, enlarging an inessential universal constant.

Lemma H.1 (Uniform localization of the projected recursion)**.** *Under Assumptions D.1, D.3 and D.4, there is a constant C depending only on (μ, L_F, γ_0, a) such that, for a localization index satisfying* (D.2)–(D.4)*, an event $\mathcal{E}_{\mathrm{loc},n}$ has probability tending to one and, on this event,*

(i) $|e_{t,j}| \le \delta_0$ *for every* $j \le k_n$ *and* $b_n \le t \le n_1$*;*

(ii) the scalar projection in (2.7) *is inactive for* $b_n < t \le n_1$*;*

(iii) uniformly over $j \le k_n$ *and* $b_n \le t \le n_1$*,*

$$|e_{t,j}| \le C\left[e^{-\frac{\mu}{2}\sum_{s=b_n+1}^{t}\gamma_s}|e_{b_n,j}| + B_{1n}\{\sqrt{\gamma_t\ell_n} + \gamma_t\ell_n\}\right]. \tag{H.2}$$

A sufficient primitive condition for selecting such a b_n is

$$B_{1n}^2\ell_n\omega_n = o(n_1^a), \qquad b_n \asymp \left(B_{1n}^2\ell_n\omega_n\right)^{1/a} + (\log n)^{1/(1-a)}, \qquad b_n = o(n_1),$$

where $\omega_n \to \infty$ may diverge arbitrarily slowly and the multiplicative constants are chosen large enough.

Proof. We use deterministic contraction weights so that every innovation is multiplied by a predictable coefficient.

Step 1: a stopped Lyapunov inequality. Projection onto $[L, U]$ is nonexpansive and fixes $Q_{j,n}$. With $h_{t-1,j} = F(q_{t-1,j}) - \tau_{j,n}$,

$$e_{t,j}^2 \le \left[e_{t-1,j} - \gamma_t\{h_{t-1,j} + \zeta_{t,j}\}\right]^2.$$

The signed secant condition gives $\mu \le h_{t-1,j}/e_{t-1,j} \le L_F$ whenever $e_{t-1,j} \neq 0$. Since $\gamma_t L_F \le 1/2$,

$$|e_{t-1,j} - \gamma_t h_{t-1,j}| \le (1-\mu\gamma_t)|e_{t-1,j}| \le |e_{t-1,j}|.$$

These inequalities also hold when $e_{t-1,j} = 0$, since then $h_{t-1,j} = 0$. Squaring the first inequality and using $(1-\mu\gamma_t)^2 \le 1-\mu\gamma_t$ gives the drift contraction. After expanding the square, define

$$\Delta M_{t,j} = -2\gamma_t(e_{t-1,j} - \gamma_t h_{t-1,j})\zeta_{t,j} + \gamma_t^2\{\zeta_{t,j}^2 - \mathbb{E}(\zeta_{t,j}^2 \mid \mathcal{F}_{t-1})\}. \tag{H.3}$$

Then $\mathbb{E}(\Delta M_{t,j} \mid \mathcal{F}_{t-1}) = 0$. The secant bounds, the gain restriction, and (H.1) give

$$e_{t,j}^2 \le (1-\mu\gamma_t)e_{t-1,j}^2 + C\gamma_t^2 B_{1n}^2 + \Delta M_{t,j}, \tag{H.4}$$

$$|\Delta M_{t,j}| \le C\gamma_t B_{1n}(|e_{t-1,j}| + \gamma_t B_{1n}), \tag{H.5}$$

$$\mathbb{E}(\Delta M_{t,j}^2 \mid \mathcal{F}_{t-1}) \le C\gamma_t^2 B_{1n}^2(e_{t-1,j}^2 + \gamma_t^2 B_{1n}^2). \tag{H.6}$$

In particular, taking expectations and iterating (H.4) yields

$$\mathbb{E}e_{t,j}^2 \le (U-L)^2 \exp\left(-\mu\sum_{s=1}^{t}\gamma_s\right) + CB_{1n}^2\gamma_t.$$

For the uniform probability bound, fix a deterministic $r < t$. Iterating (H.4) from r to t involves only the deterministic weights

$$\rho_{s,t} = \prod_{u=s+1}^{t}(1-\mu\gamma_u), \qquad \rho_{t,t} = 1.$$

Because $\mu\gamma_u \le 1/2$,

$$\rho_{s,t} \le \exp\left(-\mu\sum_{u=s+1}^{t}\gamma_u\right).$$

For every fixed $q \ge 1$ and $m > 1$,

$$\sum_{s=r+1}^{t}\rho_{s,t}^q\gamma_s^m \le C_{q,m}\gamma_t^{m-1}, \qquad \max_{r<s\le t}\rho_{s,t}^q\gamma_s^m \le C_{q,m}\gamma_t^m. \tag{H.7}$$

The maximum inequality in (H.7) also holds for every $m > 0$.

To verify these bounds, split the index set at $t/2$. If $s > t/2$, then $\gamma_s \le 2^a\gamma_t$ and

$$\rho_{s,t}^q \le \exp\{-q\mu(t-s)\gamma_t\}.$$

Consequently,

$$\sum_{t/2<s\le t}\rho_{s,t}^q\gamma_s^m \le C\gamma_t^m\sum_{h=0}^{\infty}e^{-q\mu h\gamma_t} \le C\gamma_t^{m-1}.$$

If $s \le t/2$, then

$$\rho_{s,t}^q \le \exp\left\{-q\mu\sum_{u=\lfloor t/2\rfloor+1}^{t}\gamma_u\right\} \le \exp(-ct^{1-a}),$$

and hence

$$\sum_{r<s\le t/2}\rho_{s,t}^q\gamma_s^m \le Cte^{-ct^{1-a}} \le C\gamma_t^{m-1}.$$

The same split proves the maximum bound. In particular,

$$\sum_{s=r+1}^{t}\rho_{s,t}\gamma_s^2 \le C\gamma_t, \qquad \max_{r<s\le t}\rho_{s,t}\gamma_s \le C\gamma_t.$$

For $x \ge 1$, define the first crossing time

$$\tau = \inf\left\{s \in \{r+1,\dots,n_1\} : e_{s,j}^2 > C_0\left[\rho_{r,s}e_{r,j}^2 + B_{1n}^2\{\gamma_s x + \gamma_s^2 x^2\}\right]\right\}, \qquad \inf\varnothing = n_1+1.$$

For a fixed candidate terminal time $t > r$, put

$$\widetilde{M}_{u,t,j} = \sum_{s=r+1}^{u} \mathbf{1}\{\tau \geq s\}\rho_{s,t}\Delta M_{s,j}, \qquad r \leq u \leq t.$$

Because $\{\tau \geq s\} \in \mathcal{F}_{s-1}$ and $\rho_{s,t}$ is deterministic, $(\widetilde{M}_{u,t,j})_{u=r}^{t}$ is a martingale. Its predictable quadratic variation and largest increment are

$$V_{r,t,j} = \sum_{s=r+1}^{t} \mathbb{E}\left[\{\mathbf{1}(\tau \geq s)\rho_{s,t}\Delta M_{s,j}\}^2 \mid \mathcal{F}_{s-1}\right],$$
$$L_{r,t,j} = \max_{r<s\leq t} |\mathbf{1}(\tau \geq s)\rho_{s,t}\Delta M_{s,j}|.$$

On $\{\tau \geq s\}$, if $s \geq r+2$ the absence of an earlier crossing gives the boundary at $s-1$; if $s = r+1$, the same bound follows directly from $e_{s-1,j} = e_{r,j}$. Thus, in both cases,

$$e_{s-1,j}^2 \leq C\left[\rho_{r,s-1}e_{r,j}^2 + B_{1n}^2\{\gamma_s x + \gamma_s^2 x^2\}\right].$$

Combining this bound with (H.5)–(H.6) gives

$$V_{r,t,j} \leq C\left[B_{1n}^2 e_{r,j}^2 \rho_{r,t}\gamma_t + B_{1n}^4\{\gamma_t^2 x + \gamma_t^3 x^2\}\right], \tag{H.8}$$
$$L_{r,t,j} \leq C\left[B_{1n}|e_{r,j}|\rho_{r,t}^{1/2}\gamma_t + B_{1n}^2\{\gamma_t^{3/2}x^{1/2} + \gamma_t^2 x\}\right]. \tag{H.9}$$

Here are the summations behind these two displays. The product identity and the gain restriction give

$$\rho_{s,t}\rho_{r,s-1} = \frac{\rho_{r,t}}{1-\mu\gamma_s}, \qquad (1-\mu\gamma_s)^{-1} \leq 2.$$

Therefore,

$$\sum_{s=r+1}^{t} \rho_{s,t}^2\gamma_s^2\rho_{r,s-1} \leq C\rho_{r,t}\sum_{s=r+1}^{t}\rho_{s,t}\gamma_s^2 \leq C\rho_{r,t}\gamma_t. \tag{H.10}$$
$$\sum_{s=r+1}^{t} \rho_{s,t}^2\gamma_s^3 \leq C\gamma_t^2, \qquad \sum_{s=r+1}^{t} \rho_{s,t}^2\gamma_s^4 \leq C\gamma_t^3.$$

The corresponding maximum bounds are

$$\max_{r<s\leq t} \rho_{s,t}\gamma_s\rho_{r,s-1}^{1/2} \leq C\rho_{r,t}^{1/2}\gamma_t, \tag{H.11}$$
$$\max_{r<s\leq t} \rho_{s,t}\gamma_s^{3/2} \leq C\gamma_t^{3/2}, \qquad \max_{r<s\leq t} \rho_{s,t}\gamma_s^2 \leq C\gamma_t^2. \tag{H.12}$$

Indeed, the square of the left-hand side of (H.11) is at most

$$C\rho_{r,t}\rho_{s,t}\gamma_s^2 \leq C\rho_{r,t}\gamma_t^2$$

by (H.7). Equations (H.10)–(H.12) prove (H.8)–(H.9).

Let $\overline{V}_{r,t,j}$ and $\overline{L}_{r,t,j}$ denote the respective right-hand sides of (H.8) and (H.9). They are $\mathcal{F}_r$-measurable upper bounds for the predictable variation and the increment size. Freedman's inequality now yields

$$\mathbb{P}\left\{\widetilde{M}_{t,t,j} > \sqrt{2x\overline{V}_{r,t,j}} + \frac{x\overline{L}_{r,t,j}}{3} \,\middle|\, \mathcal{F}_r\right\} \leq e^{-x}.$$

The initialization terms in the Freedman threshold satisfy, for every fixed $\eta > 0$,

$$B_{1n}|e_{r,j}|\sqrt{\rho_{r,t}\gamma_t x} \leq \eta\rho_{r,t}e_{r,j}^2 + C_\eta B_{1n}^2\gamma_t x,$$
$$xB_{1n}|e_{r,j}|\rho_{r,t}^{1/2}\gamma_t \leq \eta\rho_{r,t}e_{r,j}^2 + C_\eta B_{1n}^2\gamma_t^2 x^2.$$

Since $(\gamma_t x)^{3/2} \leq \gamma_t x + (\gamma_t x)^2$,

$$\sqrt{2x\overline{V}_{r,t,j}} + \frac{x\overline{L}_{r,t,j}}{3} \leq C\left[\rho_{r,t}e_{r,j}^2 + B_{1n}^2\{\gamma_t x + \gamma_t^2 x^2\}\right].$$

On $\{\tau = t\}$ all indicators in $\widetilde{M}_{t,t,j}$ equal one, and iteration of (H.4) gives

$$e_{t,j}^2 \le \rho_{r,t} e_{r,j}^2 + C B_{1n}^2 \gamma_t + \widetilde{M}_{t,t,j}.$$

Since $x \ge 1$, the middle term is absorbed by $C B_{1n}^2 \gamma_t x$. Choosing C_0 larger than the constants above therefore implies

$$\{\tau = t\} \subseteq \left\{ \widetilde{M}_{t,t,j} > \sqrt{2x\overline{V}_{r,t,j}} + \frac{x\overline{L}_{r,t,j}}{3} \right\}.$$

Summing its probability over the possible first crossing times proves

$$\mathbb{P}\left(\exists t \in \{r+1, \dots, n_1\} : \; e_{t,j}^2 > C\left[\rho_{r,t} e_{r,j}^2 + B_{1n}^2\{\gamma_t x + \gamma_t^2 x^2\}\right] \mid \mathcal{F}_r\right) \le 2(n_1 - r)e^{-x}. \tag{H.13}$$

This argument is the bounded-noise, one-dimensional stopped Lyapunov specialization of the projected stochastic-approximation stability proof; (H.3)–(H.13) display the martingale and its predictable weights explicitly.

Step 2: uniform entry and inactive projection. Apply (H.13) with $r = 0$ and $x = c\ell_n$, and take a union bound over $j \le k_n$. Because $\rho_{0,t} \le \exp(-\mu \sum_{s=1}^t \gamma_s)$, conditions (D.3)–(D.4) imply, for all sufficiently large n, that $|e_{t,j}| \le \delta_0$ simultaneously for $t \ge b_n$ and $j \le k_n$. If $t > b_n$, the unprojected displacement is bounded by

$$\gamma_t\{|h_{t-1,j}| + |\zeta_{t,j}|\} \le C\gamma_{b_n} B_{1n} = o(\delta_0).$$

Since every $Q_{j,n}$ is at least $3\delta_0$ from the endpoints of the public bracket, the unprojected update remains in $[L, U]$. Projection is therefore inactive after b_n.

Finally apply (H.13) conditionally with $r = b_n$. The deterministic weights satisfy

$$\rho_{b_n,t}^{1/2} \le \exp\left(-\frac{\mu}{2} \sum_{s=b_n+1}^{t} \gamma_s\right).$$

Taking square roots and union bounding over (t, j) gives (H.2) and completes the three assertions.

Step 3: a concrete b_n. For $\gamma_t = \gamma_0 t^{-a}$,

$$\sum_{s=1}^{b_n} \gamma_s \asymp b_n^{1-a}, \qquad \gamma_{b_n} \asymp b_n^{-a}.$$

The displayed order for b_n simultaneously makes the initialization factor polynomially small and $B_{1n}\sqrt{\gamma_{b_n}\ell_n} = O(\omega_n^{-1/2}) = o(1)$. The corresponding requirement $B_{1n}^2 \ell_n \omega_n = o(n_1^a)$ permits $b_n = o(n_1)$. This completes the proof. □

Lemma H.2 (Abel decomposition for the full average)**.** *Suppose Assumption D.2 holds. On $\mathcal{E}_{\mathrm{loc},n}$, for each j,*

$$\begin{aligned} f(Q_{j,n})\frac{1}{n_1}\sum_{t=b_n+1}^{n_1} e_{t-1,j} = &-\frac{1}{n_1}\sum_{t=b_n+1}^{n_1}\zeta_{t,j} - \frac{1}{n_1}\sum_{t=b_n+1}^{n_1}\mathcal{R}_{t-1,j} \\ &+ \frac{1}{n_1}\left\{\frac{e_{b_n,j}}{\gamma_{b_n+1}} - \frac{e_{n_1,j}}{\gamma_{n_1}} + \sum_{t=b_n+1}^{n_1-1} e_{t,j}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\right\}, \end{aligned} \tag{H.14}$$

where

$$\mathcal{R}_{t-1,j} = F(Q_{j,n} + e_{t-1,j}) - F(Q_{j,n}) - f(Q_{j,n})e_{t-1,j}$$

satisfies

$$|\mathcal{R}_{t-1,j}| \le \frac{L_f}{1+\beta}|e_{t-1,j}|^{1+\beta}. \tag{H.15}$$

Proof. Projection is inactive after b_n, so

$$e_{t,j} = e_{t-1,j} - \gamma_t\{F(Q_{j,n} + e_{t-1,j}) - F(Q_{j,n}) + \zeta_{t,j}\}.$$

Rearrange this equality to obtain

$$f(Q_{j,n})e_{t-1,j} = \frac{e_{t-1,j} - e_{t,j}}{\gamma_t} - \zeta_{t,j} - \mathcal{R}_{t-1,j}.$$

Summing from $b_n + 1$ to n_1 gives the first two terms in (H.14). For the difference term,

$$\sum_{t=b_n+1}^{n_1} \frac{e_{t-1,j} - e_{t,j}}{\gamma_t} = \frac{e_{b_n,j}}{\gamma_{b_n+1}} - \frac{e_{n_1,j}}{\gamma_{n_1}} + \sum_{t=b_n+1}^{n_1-1} e_{t,j}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right),$$

which is summation by parts. Finally,

$$\mathcal{R}_{t-1,j} = e_{t-1,j}\int_0^1 \{f(Q_{j,n} + se_{t-1,j}) - f(Q_{j,n})\}\, ds,$$

so Assumption D.2 yields (H.15). □

Lemma H.3 (Uniform bounds for the terms in (H.14))**.** *Under Assumptions D.1 to D.4, on events whose probabilities tend to one,*

$$\max_{j\le k_n}\left|\frac{1}{n_1}\sum_{t=b_n+1}^{n_1}\zeta_{t,j}\right| \le CB_{1n}\sqrt{\frac{\ell_n}{n_1}}, \tag{H.16}$$

$$\max_{j\le k_n}\frac{1}{n_1}\left|\frac{e_{b_n,j}}{\gamma_{b_n+1}} - \frac{e_{n_1,j}}{\gamma_{n_1}} + \sum_{t=b_n+1}^{n_1-1} e_{t,j}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\right| \le CB_{1n}n_1^{a/2-1}\sqrt{\ell_n} + o(r_{1n}), \tag{H.17}$$

$$\max_{j\le k_n}\frac{1}{n_1}\sum_{t=b_n+1}^{n_1}|\mathcal{R}_{t-1,j}| \le CB_{1n}^{1+\beta}n_1^{-a(1+\beta)/2}\ell_n^{(1+\beta)/2} + o(r_{1n}). \tag{H.18}$$

The displayed $o(r_{1n})$ terms collect the exponentially decaying initialization contribution.

Proof. For (H.16), apply Bernstein's martingale inequality to the bounded differences $\zeta_{t,j}$ and use a union bound over j. The resulting bound is

$$CB_{1n}\left\{\sqrt{\frac{\ell_n}{n_1}} + \frac{\ell_n}{n_1}\right\};$$

By (D.4), $b_n < n_1$, and $\gamma_{b_n} = \gamma_0 b_n^{-a}$, $B_{1n}^2\ell_n = o(b_n^a) = o(n_1^a)$. Since $a < 1$, this implies $\ell_n/n_1 = o(1)$, so the second term is dominated by the first.

For (H.17), write $b = b_n$ and put

$$D_{b,t} = \exp\left(-\frac{\mu}{2}\sum_{s=b+1}^{t}\gamma_s\right), \qquad D_{b,b} = 1.$$

The same near–far split used in (H.7) gives, for every fixed $q > 0$,

$$\sup_{t\ge b} D_{b,t}\left(\frac{\gamma_b}{\gamma_t}\right)^q \le C_q, \tag{H.19}$$

$$\sum_{t=b}^{N} D_{b,t}^q \le C_q\gamma_b^{-1}, \tag{H.20}$$

$$\frac{D_{b,N}}{\gamma_N} + \sum_{t=b+1}^{N-1} D_{b,t}t^{a-1} \le C\gamma_b^{-1}. \tag{H.21}$$

Indeed, when $b < t \le 2b$,

$$\sum_{s=b+1}^{t}\gamma_s \ge c(t-b)\gamma_b,$$

so the corresponding sums are dominated by a geometric series with effective length $O(\gamma_b^{-1})$. When $t > 2b$,

$$\sum_{s=b+1}^{t} \gamma_s \geq c(t^{1-a} - b^{1-a}),$$

and the remaining tail is exponentially small.

Applying (H.13) with $r = 0$ and $x = c\ell_n$ at time b gives

$$|e_{b,j}| \leq d_{0n} + CB_{1n}\{\sqrt{\gamma_b \ell_n} + \gamma_b \ell_n\}, \qquad d_{0n} = (U - L)e^{-\frac{\mu}{2}\sum_{s=1}^{b}\gamma_s} = o(r_{1n}).$$

Combining this display with (H.2) and (H.19) yields

$$|e_{t,j}| \leq CD_{b,t}d_{0n} + CB_{1n}\{\sqrt{\gamma_t \ell_n} + \gamma_t \ell_n\}. \tag{H.22}$$

Here we used, respectively, $q = 1/2$ and $q = 1$ in (H.19) to propagate the two stochastic terms at time b. By (D.4), $B_{1n} \geq 1$, and monotonicity of the gain, $\gamma_t \ell_n \leq \gamma_b \ell_n = o(1)$ for $t \geq b$; hence the second term in (H.22) is bounded by $CB_{1n}t^{-a/2}\sqrt{\ell_n}$.

The contribution of the deterministic initialization to the Abel bracket is, by (H.21),

$$\frac{Cd_{0n}}{n_1}\left[\gamma_b^{-1} + \frac{D_{b,n_1}}{\gamma_{n_1}} + \sum_{t=b+1}^{n_1-1} D_{b,t}t^{a-1}\right] \leq \frac{Cd_{0n}}{n_1\gamma_b} = o(r_{1n}),$$

because $b = o(n_1)$ and $a < 1$ imply $n_1\gamma_b \asymp n_1 b^{-a} \to \infty$. For the remaining terms,

$$\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} = \gamma_0^{-1}\{(t+1)^a - t^a\} \leq Ct^{a-1},$$

and therefore

$$\sum_{t=b+1}^{n_1-1} |e_{t,j}| \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right) \leq CB_{1n}\sqrt{\ell_n} \sum_{t \leq n_1} t^{a/2-1} + o(n_1 r_{1n}) \leq CB_{1n}\sqrt{\ell_n}\, n_1^{a/2} + o(n_1 r_{1n}).$$

The two endpoint terms obey the same bound:

$$\frac{|e_{b,j}|}{\gamma_{b+1}} + \frac{|e_{n_1,j}|}{\gamma_{n_1}} \leq CB_{1n}\sqrt{\ell_n}\, n_1^{a/2} + o(n_1 r_{1n}).$$

Dividing by n_1 proves (H.17).

For (H.18), combine (H.15) and (H.22). Since $(x+y)^{1+\beta} \leq 2^\beta(x^{1+\beta} + y^{1+\beta})$,

$$|e_{t,j}|^{1+\beta} \leq CD_{b,t}^{1+\beta}d_{0n}^{1+\beta} + CB_{1n}^{1+\beta}\gamma_t^{(1+\beta)/2}\ell_n^{(1+\beta)/2}.$$

The stochastic sum is bounded by

$$\frac{B_{1n}^{1+\beta}\ell_n^{(1+\beta)/2}}{n_1} \sum_{t \leq n_1} t^{-a(1+\beta)/2} \leq CB_{1n}^{1+\beta}\ell_n^{(1+\beta)/2} n_1^{-a(1+\beta)/2},$$

where $a(1+\beta)/2 < 1$. By (H.20), the initialization contribution is at most

$$\frac{Cd_{0n}^{1+\beta}}{n_1\gamma_b} = o(r_{1n}),$$

because $d_{0n} = o\{\min(r_{1n}, 1)\}$ and $n_1\gamma_b \to \infty$. This proves all three bounds. □

Theorem H.4 (Full-average preliminary rate)**.** *Suppose Assumptions D.1 to D.4 hold. Then the full online average in* (2.8) *satisfies*

$$\max_{j \leq k_n} |q_j^0 - Q_{j,n}| = O_{\mathbb{P}}(r_{1n}). \tag{H.23}$$

On an event with probability tending to one, it also admits the expansion

$$f(Q_{j,n})(q_j^0 - Q_{j,n}) = -\frac{1}{n_1} \sum_{t=b_n+1}^{n_1} \zeta_{t,j} + R_{1n,j}, \tag{H.24}$$

where $\zeta_{t,j} = G_{t,j}^{(1)} - F(q_{t-1,j})$ and $\max_j |R_{1n,j}|$ has the order given by the early-iterate, Abel-endpoint, and nonlinear terms in (D.5)*.*

Proof. Write $f_j = f(Q_{j,n})$. Since

$$\sum_{t=b_n+1}^{n_1} e_{t,j} = \sum_{t=b_n+1}^{n_1} e_{t-1,j} + e_{n_1,j} - e_{b_n,j},$$

Lemma H.2 gives the exact expansion

$$f_j(q_j^0 - Q_{j,n}) = -\frac{1}{n_1} \sum_{t=b_n+1}^{n_1} \zeta_{t,j} + R_{1n,j},$$

where

$$R_{1n,j} = \frac{f_j}{n_1} \sum_{t=1}^{b_n} e_{t,j} + \frac{f_j}{n_1}(e_{n_1,j} - e_{b_n,j}) - \frac{1}{n_1} \sum_{t=b_n+1}^{n_1} \mathcal{R}_{t-1,j} + \frac{1}{n_1} \left\{ \frac{e_{b_n,j}}{\gamma_{b_n+1}} - \frac{e_{n_1,j}}{\gamma_{n_1}} + \sum_{t=b_n+1}^{n_1-1} e_{t,j} \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \right\}. \quad \text{(H.25)}$$

The first term is at most $f_+(U-L)b_n/n_1$. By (H.22),

$$\frac{|e_{n_1,j}| + |e_{b_n,j}|}{n_1} \le \frac{Cd_{0n}}{n_1} + CB_{1n} n_1^{a/2-1} \sqrt{\ell_n} = o(r_{1n}) + O\Big(B_{1n} n_1^{a/2-1} \sqrt{\ell_n} \Big).$$

The last two terms in (H.25) are bounded by (H.18) and (H.17). Together with (H.16), these are exactly the four orders in (D.5). Since $f_j \ge f_-$ uniformly, this proves (H.23) and (H.24). □

Proposition H.5 (Full versus delayed averaging)**.** *For $0 \le b < n_1$, let*

$$\overline{q}_{n_1,j}^{(b)} = (n_1 - b)^{-1} \sum_{t=b+1}^{n_1} q_{t,j}.$$

Then, deterministically,

$$\max_{j \le k_n} |\overline{q}_{n_1,j}^{(b)} - \overline{q}_{n_1,j}| \le (U-L)\frac{b}{n_1}.$$

Proof. For $b > 0$,

$$\overline{q}_{n_1,j} = \frac{b}{n_1} \left(\frac{1}{b} \sum_{t=1}^{b} q_{t,j} \right) + \frac{n_1 - b}{n_1} \overline{q}_{n_1,j}^{(b)}.$$

Therefore

$$\overline{q}_{n_1,j}^{(b)} - \overline{q}_{n_1,j} = \frac{b}{n_1} \left\{ \overline{q}_{n_1,j}^{(b)} - \frac{1}{b} \sum_{t=1}^{b} q_{t,j} \right\}.$$

Both terms in braces lie in $[L, U]$, so the absolute difference is at most $(U-L)b/n_1$. The case $b = 0$ is immediate. □

I Private Density Estimation and the One-Step Expansion

Lemma I.1 (Conditional concentration of Stage-2 CDF scores)**.** *Under Assumption D.1, conditional on Stage 1, with conditional probability tending to one,*

$$\max_{j \le k_n} \left\{ \left| \overline{G}_{j,-}^{(2)} - F(q_j^0 - h_n) \right|, \left| \overline{G}_{j,+}^{(2)} - F(q_j^0 + h_n) \right| \right\} \le CB_{2n} \left\{ \sqrt{\frac{\ell_n}{n_2}} + \frac{\ell_n}{n_2} \right\}.$$

Proof. Once Stage 1 is fixed, the $2k_n$ thresholds, their ranks, and all anchors are deterministic. The Stage-2 user vectors are conditionally i.i.d. By Proposition 2.1, each coordinate is unbiased and bounded in absolute value by B_{2n}. Therefore the centered coordinate is bounded by $2B_{2n}$. For either sign, Bernstein's inequality gives, for any $x > 0$,

$$\mathbb{P}\left(|\overline{G}_{j,\pm}^{(2)} - F(q_j^0 \pm h_n)| > CB_{2n}\{\sqrt{x/n_2} + x/n_2\} \mid \mathcal{F}_{n_1} \right) \le 2e^{-x}.$$

Set $x = c\ell_n$ and union bound over $2k_n$ tagged coordinates. Integration over Stage 1 proves the result. □

Theorem I.2 (Uniform private density rate)**.** *Suppose Assumptions D.1, D.2 and D.5 hold and* (H.23) *is valid. Then*

$$\max_{j\le k_n} |\widehat{f}_j - f(Q_{j,n})| = O_{\mathbb{P}}(d_n). \tag{I.1}$$

With probability tending to one, all $q_j^0 \pm h_n$ lie in the interior neighborhood and both clipping operations in (2.10) *are inactive.*

Proof. By Theorem H.4, $\max_j |q_j^0 - Q_{j,n}| = O_{\mathbb{P}}(r_{1n}) = o_{\mathbb{P}}(\delta_0)$, while $h_n = o(\delta_0)$. Hence, with probability tending to one, $q_j^0 \pm h_n$ lie in $\mathcal{N}_n$ and in the interior of $[L, U]$ for every j; endpoint projection is inactive.

Define the conditional population finite difference

$$f_{j,h}^0 = \frac{F(q_j^0 + h_n) - F(q_j^0 - h_n)}{2h_n} = \frac{1}{2h_n}\int_{-h_n}^{h_n} f(q_j^0 + u)\, du.$$

For every $u \in [-h_n, h_n]$, Assumption D.2 gives

$$\begin{aligned} |f(q_j^0 + u) - f(Q_{j,n})| &\le L_f |q_j^0 - Q_{j,n} + u|^\beta \\ &\le L_f\{|q_j^0 - Q_{j,n}|^\beta + |u|^\beta\}, \end{aligned}$$

where $(x+y)^\beta \le x^\beta + y^\beta$ for $\beta \le 1$. Averaging over u yields

$$\max_j |f_{j,h}^0 - f(Q_{j,n})| \le C\left\{\max_{\ell\le k_n} |q_\ell^0 - Q_{\ell,n}|^\beta + h_n^\beta\right\} = O_{\mathbb{P}}\{r_{1n}^\beta + h_n^\beta\}. \tag{I.2}$$

By Lemma I.1,

$$\begin{aligned} \max_j |\widetilde{f}_j - f_{j,h}^0| &\le \frac{1}{2h_n}\max_j \left(|\overline{G}_{j,+}^{(2)} - F(q_j^0 + h_n)| + |\overline{G}_{j,-}^{(2)} - F(q_j^0 - h_n)|\right) \\ &\le C\frac{B_{2n}}{h_n}\left\{\sqrt{\frac{\ell_n}{n_2}} + \frac{\ell_n}{n_2}\right\}. \end{aligned} \tag{I.3}$$

Combining (I.2) and (I.3) gives the rate d_n for $\widetilde{f}_j$.

Because $d_n \to 0$, $f(Q_{j,n}) \in [f_-, f_+]$, $\varpi_n \to 0$, and $\varpi_n^{-1} \to \infty$, eventually

$$\varpi_n < f_-/2, \qquad \varpi_n^{-1} > 2f_+,$$

and the event $\max_j |\widetilde{f}_j - f(Q_{j,n})| < f_-/2$ has probability tending to one. On that event,

$$\frac{f_-}{2} \le \widetilde{f}_j \le f_+ + \frac{f_-}{2} \le \frac{3f_+}{2} < 2f_+ < \varpi_n^{-1},$$

while $\varpi_n < f_-/2$. Hence $\varpi_n < \widetilde{f}_j < \varpi_n^{-1}$ for every j, so neither clipping boundary is active and $\widehat{f}_j = \widetilde{f}_j$. This proves (I.1). □

Proof of Theorem 3.1. Conditional on the first two stages, write

$$\overline{G}_j^{(3)} = F(q_j^0) + \overline{Y}_j, \qquad \overline{Y}_j = \frac{1}{n_3}\sum_{i=1}^{n_3} Y_{i,j}.$$

Substituting this identity into (2.11) and multiplying by $\widehat{f}_j$ gives the exact equality

$$\widehat{f}_j(\check{Q}_j - Q_{j,n}) + \overline{Y}_j = \widehat{f}_j(q_j^0 - Q_{j,n}) - \{F(q_j^0) - F(Q_{j,n})\}. \tag{I.4}$$

Add and subtract $f(Q_{j,n})(q_j^0 - Q_{j,n})$ on the right:

$$\begin{aligned} \text{(I.4)} = \{\widehat{f}_j - f(Q_{j,n})\}(q_j^0 - Q_{j,n}) \\ - \left[F(q_j^0) - F(Q_{j,n}) - f(Q_{j,n})(q_j^0 - Q_{j,n})\right]. \end{aligned} \tag{I.5}$$

The first term is uniformly $O_{\mathbb{P}}(d_n r_{1n})$ by Theorems H.4 and I.2. For the second, the integral argument used in (H.15) gives

$$\left|F(q_j^0) - F(Q_{j,n}) - f(Q_{j,n})(q_j^0 - Q_{j,n})\right| \le \frac{L_f}{1+\beta}|q_j^0 - Q_{j,n}|^{1+\beta},$$

which is uniformly $O_{\mathbb{P}}(r_{1n}^{1+\beta})$. Taking the maximum proves (3.1). □

J High-Dimensional Gaussian and Multiplier Approximation

Lemma J.1 (Conditional variance and standardized envelope)**.** *Suppose Assumptions D.1 and D.2 hold and*

$$\max_{j\le k_n} |q_j^0 - Q_{j,n}| = o_{\mathbb{P}}(1).$$

Then there exists a deterministic sequence $\eta_n \downarrow 0$ such that, with

$$\mathcal{E}_{0,n} = \left\{ \max_{j\le k_n} |q_j^0 - Q_{j,n}| \le \eta_n \right\}, \qquad \mathbb{P}(\mathcal{E}_{0,n}) \longrightarrow 1,$$

there are constants $0 < c < C < \infty$, depending only on (τ_-, τ_+), for which the following bounds hold on $\mathcal{E}_{0,n}$ for all sufficiently large n:

$$cH_{k_n}(1+\nu_{\varepsilon_n}) \le \min_{j\le k_n} \sigma_{j,n}^2 \le \max_{j\le k_n} \sigma_{j,n}^2 \le CH_{k_n}^2(1+\nu_{\varepsilon_n}).$$

Consequently, on the same event,

$$K_n \le C\sqrt{H_{k_n}}. \tag{J.1}$$

Proof. Choose $\eta_n < \delta_0$ eventually. On $\mathcal{E}_{0,n}$, the segment joining $Q_{j,n}$ and q_j^0 lies in $\mathcal{N}_n$, and hence

$$|p_j^0 - \tau_{j,n}| = |F(q_j^0) - F(Q_{j,n})| \le f_+\eta_n.$$

Using the target-range part of Assumption D.1, for all sufficiently large n,

$$p_j^0 = F(q_j^0) \in [\tau_-/2, (1+\tau_+)/2]$$

uniformly in j. Thus both p_j^0 and $1-p_j^0$ are bounded below by a positive constant.

For a threshold rank r, define $a = a_r^-(p_j^0)$ and $b = a_r^+(p_j^0)$. Since $1 \le w_r^{\pm} \le H_{k_n}$ and $2H_{k_n} - 1 \ge H_{k_n}$,

$$a \ge H_{k_n}\nu_{\varepsilon_n} + cH_{k_n}, \qquad b \ge H_{k_n}\nu_{\varepsilon_n} + cH_{k_n}.$$

The minimum over all fusion weights equals

$$p_j^0(1-p_j^0) + \frac{ab}{a+b} \ge c + \frac{1}{2}\min(a,b) \ge cH_{k_n}(1+\nu_{\varepsilon_n}).$$

The variance at the actual anchor τ_j is no smaller than this minimum, proving the lower bound.

For the upper bound, a and b are at most $CH_{k_n}^2(1+\nu_{\varepsilon_n})$, and (F.4) at any $\lambda \in [0,1]$ is bounded by the same order. Alternatively, the score envelope gives a coarser bound of the same required form. Finally,

$$1 + \nu_\varepsilon = 1 + \frac{4e^\varepsilon}{(e^\varepsilon - 1)^2} = \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1}\right)^2,$$

so

$$B_{1n} = 1 + \frac{2H_{k_n}e^{\varepsilon_n}}{e^{\varepsilon_n} - 1} \le 3H_{k_n}\sqrt{1+\nu_{\varepsilon_n}}.$$

Therefore

$$\frac{|Y_{i,j}|}{\sigma_{j,n}} \le \frac{2B_{1n}}{\{cH_{k_n}(1+\nu_{\varepsilon_n})\}^{1/2}} \le C\sqrt{H_{k_n}},$$

which proves (J.1). □

Lemma J.2 (Uniform sample-scale consistency)**.** *Conditional on the first two stages,*

$$\max_{j\le k_n} \left| \frac{\widehat{\sigma}_{G,j}^2}{\sigma_{j,n}^2} - 1 \right| = O_{\mathbb{P}}\left[K_n\sqrt{\frac{\ell_n}{n_3}} + K_n^2\frac{\ell_n}{n_3} \right]. \tag{J.2}$$

Under (D.6)*, the right-hand side is $o_{\mathbb{P}}(1)$.*

Proof. Let $X_{i,j} = Y_{i,j}/\sigma_{j,n}$ and $X_i = (X_{i,1}, \dots, X_{i,k_n})^\top$. Conditional on the first two stages, the vectors are i.i.d., $\mathbb{E}X_{i,j} = 0$, $\mathbb{E}X_{i,j}^2 = 1$, and $|X_{i,j}| \le K_n$. In particular, $\mathbb{E}X_{i,j}^4 \le K_n^2$. Bernstein's inequality and a union bound give

$$\max_j \left| n_3^{-1} \sum_i X_{i,j}^2 - 1 \right| = O_{\mathbb{P}} \left\{ K_n \sqrt{\ell_n / n_3} + K_n^2 \ell_n / n_3 \right\}.$$

The squared sample means satisfy

$$\max_j \left| n_3^{-1} \sum_i X_{i,j} \right|^2 = O_{\mathbb{P}}(K_n^2 \ell_n / n_3),$$

again by Bernstein. Since

$$\frac{\widehat{\sigma}_{G,j}^2}{\sigma_{j,n}^2} = n_3^{-1} \sum_i X_{i,j}^2 - \left(n_3^{-1} \sum_i X_{i,j} \right)^2,$$

the result follows. □

Lemma J.3 (Conditional Gaussian and empirical multiplier approximation)**.** *Let*

$$S_n^\dagger = \max_{j \in \widehat{\mathcal{J}}_n} \left| \frac{n_3^{-1/2} \sum_{i=1}^{n_3} Y_{i,j}}{\widehat{\sigma}_{G,j}} \right|, \qquad \Delta_{\sigma,n} = K_n \sqrt{\frac{\ell_n}{n_3}} + K_n^2 \frac{\ell_n}{n_3},$$

and let S_n^* be the empirical Gaussian multiplier analogue centered by the sample means and divided by the sample standard deviations. Under (D.6)*, conditional on the first two stages,*

$$\begin{aligned} &\sup_z \left| \mathbb{P}(S_n^\dagger \le z \mid \mathcal{F}_{n_1+n_2}) - \mathbb{P}(S_n^* \le z \mid G_1^{(3)}, \dots, G_{n_3}^{(3)}) \right| \\ &\quad = O_{\mathbb{P}} \left\{ \left(\frac{K_n^2 \ell_n^7}{n_3} \right)^{1/6} + \ell_n \Delta_{\sigma,n} \right\} = o_{\mathbb{P}}(1). \end{aligned} \tag{J.3}$$

Proof. Work conditionally on $\mathcal{F}_{n_1+n_2}$. Put

$$X_{i,j} = \frac{Y_{i,j}}{\sigma_{j,n}}, \qquad p_n = \max\{2k_n, 3\}.$$

Define the population-standardized target and multiplier maxima

$$\begin{aligned} S_n^0 &= \max_{j \le k_n} \left| n_3^{-1/2} \sum_{i=1}^{n_3} X_{i,j} \right|, \\ S_n^{*,0} &= \max_{j \le k_n} \left| n_3^{-1/2} \sum_{i=1}^{n_3} \xi_i \left(X_{i,j} - n_3^{-1} \sum_{\ell=1}^{n_3} X_{\ell,j} \right) \right|. \end{aligned}$$

Let Z_n be the maximum of the absolute coordinates of the centered Gaussian vector with the covariance matrix of $n_3^{-1/2} \sum_i X_i$, and let F_n^G be its conditional CDF.

Embed the absolute maximum into $(X_i^\top, -X_i^\top)^\top$ and, if $2k_n < 3$, append a duplicate coordinate. Every coordinate is centered, has variance one, and is bounded by K_n. Furthermore,

$$\log(p_n n_3) \le C \ell_n,$$

and, when $k_n \ge 2$, $\log(p_n n_3) \le \ell_n$. For $m = 1, 2$,

$$\mathbb{E}(|X_{i,j}|^{2+m} \mid \mathcal{F}_{n_1+n_2}) \le K_n^m \mathbb{E}(X_{i,j}^2 \mid \mathcal{F}_{n_1+n_2}) = K_n^m.$$

Also, after replacing K_n by a universal multiple,

$$\mathbb{E}\{\exp(|X_{i,j}|/(CK_n)) \mid \mathcal{F}_{n_1+n_2}\} \le 2.$$

Hence Proposition 2.1 of Chernozhukov et al. (2017) gives

$$\sup_z \left|\mathbb{P}(S_n^0 \le z \mid \mathcal{F}_{n_1+n_2}) - F_n^G(z)\right| \le C\left(\frac{K_n^2\ell_n^7}{n_3}\right)^{1/6}. \tag{J.4}$$

Put $m = n_3$ and $\rho_n = K_n^2\ell_n^7/m$. For the multiplier comparison, apply Corollary 4.2 of Chernozhukov et al. (2017) with $\eta = (m+1)^{-2}$. Conditional on the first two stages, with probability at least $1-(m+1)^{-2}$ over the Stage-3 sample,

$$\sup_z \left|\mathbb{P}(S_n^{*,0} \le z \mid X_1, \dots, X_m) - F_n^G(z)\right| \le C\rho_n^{1/6}, \tag{J.5}$$

because $\log^5(p_n m)\log^2\{(m+1)^2\} \le C\ell_n^7$.

It remains to replace the population scales by sample scales. Write

$$\delta_{\sigma,n} = \max_j \left|\frac{\widehat{\sigma}_{G,j}^2}{\sigma_{j,n}^2} - 1\right|, \qquad \overline{X}_j = m^{-1}\sum_{i=1}^m X_{i,j}.$$

For each coordinate, $|X_{i,j}| \le K_n$, $\mathbb{E}(X_{i,j}^2 \mid \mathcal{F}_{n_1+n_2}) = 1$, and $\mathbb{E}(X_{i,j}^4 \mid \mathcal{F}_{n_1+n_2}) \le K_n^2$. Conditional Bernstein inequalities and a union bound therefore give

$$\mathbb{P}\left[\max_j \left|m^{-1}\sum_{i=1}^m (X_{i,j}^2 - 1)\right| > C\Delta_{\sigma,n} \;\middle|\; \mathcal{F}_{n_1+n_2}\right] \le Ce^{-c\ell_n}, \tag{J.6}$$

$$\mathbb{P}\left[\max_j |\overline{X}_j| > C\left\{\sqrt{\frac{\ell_n}{m}} + K_n\frac{\ell_n}{m}\right\} \;\middle|\; \mathcal{F}_{n_1+n_2}\right] \le Ce^{-c\ell_n}. \tag{J.7}$$

The constants in the thresholds can be enlarged so that c is arbitrarily large. On first-two-stage histories with ρ_n smaller than a fixed constant, the square of the threshold in (J.7) is bounded by $C\Delta_{\sigma,n}$. Since

$$\frac{\widehat{\sigma}_{G,j}^2}{\sigma_{j,n}^2} = m^{-1}\sum_{i=1}^m X_{i,j}^2 - \overline{X}_j^2,$$

(J.6)–(J.7) imply

$$\mathbb{P}(\delta_{\sigma,n} > C\Delta_{\sigma,n} \mid \mathcal{F}_{n_1+n_2}) \le Ce^{-c\ell_n}. \tag{J.8}$$

Moreover, $\rho_n = o_{\mathbb{P}}(1)$ by (D.6), and on these histories

$$\Delta_{\sigma,n} = \rho_n^{1/2}\ell_n^{-3} + \rho_n\ell_n^{-6} = o(1).$$

Thus (J.8) also ensures $\delta_{\sigma,n} \le 1/2$ with the stated probability.

On $\{\delta_{\sigma,n} \le 1/2\}$ all sample scales are positive and

$$|S_n^\dagger - S_n^0| \le C\delta_{\sigma,n}S_n^0,$$
$$|S_n^* - S_n^{*,0}| \le C\delta_{\sigma,n}S_n^{*,0}.$$

Another conditional Bernstein inequality and union bound give

$$\mathbb{P}(S_n^0 > C\sqrt{\ell_n} \mid \mathcal{F}_{n_1+n_2}) \le Ce^{-c\ell_n}.$$

On the sample-scale event, every coordinate of the conditional Gaussian vector defining $S_n^{*,0}$ has variance at most $3/2$, so a Gaussian union bound gives

$$\mathbb{P}(S_n^{*,0} > C\sqrt{\ell_n} \mid X_1, \dots, X_m) \le Ce^{-c\ell_n}.$$

Consequently, with $t_n = C\Delta_{\sigma,n}\sqrt{\ell_n}$, the two studentization perturbations exceed t_n only on the explicit exceptional events above. Nazarov's inequality for the signed Gaussian vector gives

$$\sup_z\{F_n^G(z+t_n) - F_n^G(z-t_n)\} \le C\ell_n\Delta_{\sigma,n}.$$

The exponentially small terms can be absorbed into $C\rho_n^{1/6}$: $K_n \geq 1$, $m \leq n$, and the exponent c can be chosen sufficiently large. Combining the perturbation sandwiches separately with (J.4) and (J.5) gives

$$\sup_z \left|\mathbb{P}(S_n^\dagger \leq z \mid \mathcal{F}_{n_1+n_2}) - F_n^G(z)\right| = O_{\mathbb{P}}\left\{\left(\frac{K_n^2\ell_n^7}{n_3}\right)^{1/6} + \ell_n\Delta_{\sigma,n}\right\}, \tag{J.9}$$

$$\sup_z \left|\mathbb{P}(S_n^* \leq z \mid G_1^{(3)}, \ldots, G_{n_3}^{(3)}) - F_n^G(z)\right| = O_{\mathbb{P}}\left\{\left(\frac{K_n^2\ell_n^7}{n_3}\right)^{1/6} + \ell_n\Delta_{\sigma,n}\right\} \tag{J.10}$$

on a set of first-two-stage histories whose probability tends to one; the second display also holds outside a Stage-3 event of conditional probability at most $Ce^{-c\ell_n} + (m+1)^{-2}$. The triangle inequality then proves (J.3).

Finally,

$$\ell_n\Delta_{\sigma,n} = \left(\frac{K_n^2\ell_n^7}{n_3}\right)^{1/2}\ell_n^{-2} + \frac{K_n^2\ell_n^7}{n_3}\ell_n^{-5} = o_{\mathbb{P}}(1),$$

which completes the proof. □

Lemma J.4 (Transfer of the one-step remainder)**.** *Suppose* (3.1)*,* (J.2)*,* $\Delta_{\sigma,n} = o_{\mathbb{P}}(1)$*, and* (D.7) *hold. Then*

$$|T_n - S_n^\dagger| = O_{\mathbb{P}}\left(\frac{\sqrt{n_3}R_n}{\underline{\sigma}_n}\right) = o_{\mathbb{P}}(\ell_n^{-1/2}). \tag{J.11}$$

Proof. By Theorem 3.1, uniformly in j,

$$\left|\sqrt{n_3}\widehat{f}_j(\check{Q}_j - Q_{j,n}) + n_3^{-1/2}\sum_i Y_{i,j}\right| = O_{\mathbb{P}}(\sqrt{n_3}R_n).$$

On the event

$$\max_j \left|\frac{\widehat{\sigma}_{G,j}^2}{\sigma_{j,n}^2} - 1\right| \leq \frac{1}{2},$$

which has probability tending to one by Lemma J.2, $\widehat{\mathcal{J}}_n = \{1, \ldots, k_n\}$ and $\min_j \widehat{\sigma}_{G,j} \geq \underline{\sigma}_n/\sqrt{2}$. Divide by the sample scales, maximize, and use the reverse triangle inequality for maxima. Condition (D.7) then proves (J.11). □

Proof of Theorem 3.2. Condition on $\mathcal{F}_{n_1+n_2}$. Let Z_n be the maximum of the absolute coordinates $Z_{n,j}$ of the centered Gaussian vector used in Lemma J.3, let F_n^G be its conditional CDF, and put

$$q_n^G(u) = \inf\{z : F_n^G(z) \geq u\}.$$

The two comparisons in (J.9) and (J.10), together with $\mathbb{P}(\widehat{\mathcal{J}}_n = \{1, \ldots, k_n\}) \to 1$, imply that there are a deterministic sequence $\delta_{G,n} \downarrow 0$ and events $\mathcal{A}_n$ with $\mathbb{P}(\mathcal{A}_n^c) = o(1)$ such that, on $\mathcal{A}_n$,

$$\sup_z \left|\mathbb{P}(S_n^\dagger \leq z \mid \mathcal{F}_{n_1+n_2}) - F_n^G(z)\right| \leq \delta_{G,n}, \tag{J.12}$$

$$\sup_z \left|\mathbb{P}(S_n^* \leq z \mid G_1^{(3)}, \ldots, G_{n_3}^{(3)}) - F_n^G(z)\right| \leq \delta_{G,n}, \tag{J.13}$$

and all sample variances are positive. This event formulation follows directly from the two $O_{\mathbb{P}}$ bounds because their common rate tends to zero.

Every coordinate of the Gaussian vector has variance one. Consequently, for every z,

$$\mathbb{P}(Z_n = z \mid \mathcal{F}_{n_1+n_2}) \leq \sum_{j=1}^{k_n}\mathbb{P}(|Z_{n,j}| = z \mid \mathcal{F}_{n_1+n_2}) = 0.$$

Thus F_n^G is continuous even if its correlation matrix is singular, and $F_n^G\{q_n^G(u)\} = u$ for $u \in (0,1)$. Write $u = 1-\alpha$ and $c_u^* = c_{1-\alpha}^*$. The second inequality in (J.13) and the definition of a generalized quantile give, on $\mathcal{A}_n$,

$$q_n^G(u - \delta_{G,n}) \leq c_u^* \leq q_n^G(u + \delta_{G,n}). \tag{J.14}$$

Indeed, at a point below the left endpoint the conditional multiplier CDF is strictly below u, whereas at the right endpoint it is at least u.

The sandwich (J.14) implies the pathwise inclusions

$$\mathcal{A}_n \cap \{S_n^\dagger \le q_n^G(u-\delta_{G,n})\} \subseteq \{S_n^\dagger \le c_u^*\} \subseteq \mathcal{A}_n^c \cup \{S_n^\dagger \le q_n^G(u+\delta_{G,n})\}.$$

Using (J.12) at the two deterministic, $\mathcal{F}_{n_1+n_2}$-measurable Gaussian quantiles and then removing $\mathcal{A}_n$ yields, uniformly for $u \in [\alpha_0, 1-\alpha_0]$,

$$\left|\mathbb{P}(S_n^\dagger \le c_u^*) - u\right| \le 2\delta_{G,n} + \mathbb{P}(\mathcal{A}_n^c) = o(1).$$

By Lemma J.4, choose a deterministic $a_n = o(\ell_n^{-1/2})$ such that, for $\mathcal{D}_n = \{|T_n - S_n^\dagger| \le a_n\}$, $\mathbb{P}(\mathcal{D}_n^c) = o(1)$. On $\mathcal{A}_n \cap \mathcal{D}_n$,

$$\{S_n^\dagger \le q_n^G(u-\delta_{G,n}) - a_n\} \subseteq \{T_n \le c_u^*\},$$
$$\{T_n \le c_u^*\} \subseteq \{S_n^\dagger \le q_n^G(u+\delta_{G,n}) + a_n\}.$$

Nazarov's inequality, applied to the $2k_n$-dimensional signed Gaussian vector, gives

$$\sup_z \{F_n^G(z+a_n) - F_n^G(z-a_n)\} \le C a_n \sqrt{\ell_n} = o(1).$$

Combining this bound with (J.12) proves (3.2). On the event that all sample variances are positive, $T_n \le c_{1-\alpha}^*$ is algebraically equivalent to simultaneous inclusion in the untruncated intervals $\mathcal{I}_{n,j}^*$. Intersecting with $[L, U]$ preserves inclusion because $Q_{j,n} \in [L, U]$. The zero-variance fallback is $[L, U]$ and its probability tends to zero, which proves (3.3).

For the Monte Carlo version, let F_n^* be the exact conditional multiplier CDF and $\widehat{F}_{n,M}^*$ its empirical CDF based on M_n independent draws. Also write

$$q_n^*(v) = \inf\{z : F_n^*(z) \ge v\}.$$

Conditionally on the data, the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Massart, 1990) gives

$$\mathbb{P}\left\{\sup_z |\widehat{F}_{n,M}^*(z) - F_n^*(z)| > \eta \;\middle|\; G_1^{(3)}, \dots, G_{n_3}^{(3)}\right\} \le 2e^{-2M_n\eta^2}. \tag{J.15}$$

For $u \in [\alpha_0, 1-\alpha_0]$ and all sufficiently large M_n, the cap in $r_{M_n,\alpha}$ is inactive. At $\widehat{c}_u^* = T^*_{(r_{M_n,\alpha})}$,

$$\widehat{F}_{n,M}^*(\widehat{c}_u^*) \ge u, \qquad \widehat{F}_{n,M}^*(\widehat{c}_u^*-) < u + M_n^{-1}.$$

On the DKW event in (J.15), these two inequalities imply

$$q_n^*(u-\eta) \le \widehat{c}_u^* \le q_n^*(u+\eta+M_n^{-1}). \tag{J.16}$$

For the lower bound, a value below $q_n^*(u-\eta)$ would have empirical CDF strictly below u; for the upper bound, a value above $q_n^*(u+\eta+M_n^{-1})$ would make the empirical left limit at the latter quantile at least $u + M_n^{-1}$, contradicting the preceding display. Take $\eta = M_n^{-1/4}$ in (J.15). Then the DKW failure probability tends to zero, both shifted levels remain in a fixed compact subset of $(0,1)$, and (J.16) together with the exact-quantile result proves the two finite-M_n assertions uniformly in α.

Now fix M and α, and abbreviate $r = r_{M,\alpha}$. On the event of positive sample variances, F_n^* is continuous because every studentized multiplier coordinate is a nondegenerate Gaussian variable. The preceding Gaussian comparisons and the remainder transfer imply

$$\sup_z \left|\mathbb{P}(T_n \le z \mid \mathcal{F}_{n_1+n_2}) - F_n^G(z)\right| = o_{\mathbb{P}}(1), \qquad \sup_z |F_n^*(z) - F_n^G(z)| = o_{\mathbb{P}}(1).$$

Since F_n^G is continuous, the conditional probability-integral transform gives

$$U_n := F_n^*(T_n) \rightsquigarrow U, \qquad U \sim \mathrm{Uniform}(0,1).$$

Conditional on the data, $T_n \le T^*_{(r)}$ exactly when at most $r-1$ of the M multiplier maxima are below T_n. Therefore

$$\mathbb{P}(T_n \le T^*_{(r)} \mid G_1^{(3)}, \dots, G_{n_3}^{(3)}) = g_{M,r}(U_n),$$

where

$$g_{M,r}(u) = \sum_{s=0}^{r-1} \binom{M}{s} u^s (1-u)^{M-s}.$$

This is a bounded continuous polynomial, and the beta integral gives

$$\int_0^1 g_{M,r}(u)\, du = \sum_{s=0}^{r-1} \binom{M}{s} \frac{s!(M-s)!}{(M+1)!} = \frac{r}{M+1}.$$

Taking expectations proves (3.4). The interval equivalence and the vanishing zero-variance fallback then give the same limit for simultaneous coverage. □

Proof of Corollary 3.3. Conditional on the Stage-3 data, every nondegenerate coordinate of the studentized Gaussian multiplier vector has variance one. Hence

$$\mathbb{P}(S_n^* > x \mid G_1^{(3)}, \dots, G_{n_3}^{(3)}) \le 2k_n e^{-x^2/2}.$$

It follows that the exact conditional critical value satisfies

$$c_{1-\alpha}^* \le \sqrt{2\log(2k_n/\alpha)} = O\{\sqrt{\log(2k_n)}\}$$

uniformly for α bounded away from zero. For the empirical critical value, (J.16) and $M_n \to \infty$ give the same order in probability.

By Lemmas J.1 and J.2,

$$\max_j \widehat{\sigma}_{G,j} = O_{\mathbb{P}}\{H_{k_n}\sqrt{1+\nu_{\varepsilon_n}}\}.$$

By Theorem I.2, $\min_j \widehat{f}_j \ge f_-/2$ with probability tending to one. Substitution in (2.14) proves (3.5). For $0 < \varepsilon_n \le 1$,

$$\sqrt{1+\nu_{\varepsilon_n}} = \frac{e^{\varepsilon_n}+1}{e^{\varepsilon_n}-1} \le \frac{C}{\varepsilon_n},$$

which proves (3.6).

It remains to verify the transparent regime. Write

$$H = H_{k_n}.$$

Since

$$H_{2k_n} = \lceil \log_2(2k_n+1) \rceil \le 1 + \lceil \log_2(k_n+1) \rceil = H+1 \le 2H,$$

and ε_n is bounded away from zero,

$$B_{1n} \asymp B_{2n} \asymp H.$$

Moreover, $H \le C\ell_n$. By Lemma J.1,

$$\underline{\sigma}_n \gtrsim H^{1/2}, \qquad K_n^2 = O(H).$$

Under $a = 2/3$, $\beta = 1$, and fixed positive stage fractions, (D.5) gives

$$r_{1n} = O\left(\frac{b_n}{n} + H\sqrt{\frac{\ell_n}{n}} + Hn^{-2/3}\sqrt{\ell_n} + H^2 n^{-2/3}\ell_n\right).$$

Using $b_n/n = o(\ell_n^{3/2} n^{-1/2})$ and $H \le C\ell_n$,

$$r_{1n} = O\left(\ell_n^{3/2} n^{-1/2} + \ell_n^3 n^{-2/3}\right).$$

The ratio of the second term to the first is

$$\frac{\ell_n^3 n^{-2/3}}{\ell_n^{3/2} n^{-1/2}} = \ell_n^{3/2} n^{-1/6} = \left(\frac{\ell_n^9}{n}\right)^{1/6} = o(1).$$

Therefore,

$$r_{1n} = O(\ell_n^{3/2} n^{-1/2}). \tag{J.17}$$

The proposed bandwidth obeys

$$h_n \asymp \left(\frac{H^2\ell_n}{n}\right)^{1/4} = H^{1/2}\ell_n^{1/4}n^{-1/4}.$$

Its leading stochastic density term is

$$\frac{B_{2n}}{h_n}\sqrt{\frac{\ell_n}{n}} \asymp h_n,$$

whereas

$$\frac{B_{2n}}{h_n}\frac{\ell_n}{n} = O\left(h_n\sqrt{\frac{\ell_n}{n}}\right) = o(h_n).$$

Also,

$$\frac{r_{1n}}{h_n} = O\left(\frac{\ell_n^{5/4}}{H^{1/2}n^{1/4}}\right) = o(1),$$

because its fourth power is bounded by $C\ell_n^5/n = o(1)$. Consequently,

$$d_n = O\left(H^{1/2}\ell_n^{1/4}n^{-1/4}\right). \tag{J.18}$$

Combining (J.17) and (J.18) gives

$$r_{1n}d_n = O\left(H^{1/2}\ell_n^{7/4}n^{-3/4}\right).$$

Since $d_n \geq h_n$ and $r_{1n} = o(h_n)$, we have $r_{1n} = o(d_n)$ and hence $r_{1n}^2 = o(r_{1n}d_n)$. Therefore

$$R_n = O\left(H^{1/2}\ell_n^{7/4}n^{-3/4}\right).$$

It follows that

$$\frac{\sqrt{n_3\ell_n}}{\underline{\sigma}_n}R_n = O(\ell_n^{9/4}n^{-1/4}) = O\left\{\left(\frac{\ell_n^9}{n}\right)^{1/4}\right\} = o(1).$$

The Gaussian approximation condition is

$$\frac{K_n^2\ell_n^7}{n_3} = O\left(\frac{H\ell_n^7}{n}\right) = O\left(\frac{\ell_n^8}{n}\right) = o(1).$$

Finally, choose $\omega_n = \ell_n$ in the primitive construction of Lemma H.1. Then

$$B_{1n}^2\ell_n\omega_n = O(H^2\ell_n^2) \leq O(\ell_n^4) = o(n^{2/3}),$$

and

$$\begin{aligned} b_n &= O\left\{(H^2\ell_n^2)^{3/2} + (\log n)^3\right\} \\ &= O\{H^3\ell_n^3 + (\log n)^3\} = O(\ell_n^6). \end{aligned}$$

Moreover,

$$\frac{b_n/n}{\ell_n^{3/2}n^{-1/2}} = O\left(\frac{\ell_n^{9/2}}{\sqrt{n}}\right) = O\left\{\left(\frac{\ell_n^9}{n}\right)^{1/2}\right\} = o(1).$$

Thus the primitive construction satisfies all three localization conditions and the additional transient requirement in the corollary.

Because $\ell_n = \log\{2k_n(n+1)\}$, condition (3.7) is equivalent to

$$\log k_n = o(n^{1/9}).$$

It therefore includes fixed grids, arbitrarily slowly growing grids, all polynomial grids, and more generally $k_n = \exp\{o(n^{1/9})\}$. □

K Proofs for Post-Processing and Fixed-Chain Comparisons

Proposition K.1 (Coverage-preserving monotone refinement)**.** *If $Q_j \in [L_j^I, U_j^I]$ for every j, then $Q_j \in [\widetilde{L}_j^I, \widetilde{U}_j^I]$ for every j.*

Proof. Fix j. On the original simultaneous coverage event, for every $\ell \le j$,

$$L_\ell^I \le Q_\ell \le Q_j,$$

where the second inequality follows from monotonicity of the quantile function. Therefore $\widetilde{L}_j^I = \max_{\ell \le j} L_\ell^I \le Q_j$. Similarly, for every $\ell \ge j$,

$$Q_j \le Q_\ell \le U_\ell^I,$$

so $Q_j \le \min_{\ell \ge j} U_\ell^I = \widetilde{U}_j^I$. Both inequalities hold for every j on the same event. □

Proof of Proposition C.1. Let

$$s_J^2 = (J-1)^{-1} \sum_{b=1}^{J} (W_b - \overline{W})^2.$$

For a normal sample, $\sqrt{J}(\overline{W} - \theta)/s_J \sim t_{J-1}$. Because

$$S_J^2 = \frac{J-1}{J} s_J^2,$$

we have

$$\frac{\sqrt{J-1}(\overline{W} - \theta)}{S_J} = \frac{\sqrt{J}(\overline{W} - \theta)}{s_J} \sim t_{J-1}.$$

Conditional on $W_1, \ldots, W_J$, the multiplier numerator is centered normal with variance

$$\mathrm{Var}\left(J^{-1/2} \sum_{b=1}^{J} \xi_b (W_b - \overline{W}) \mid W_1, \ldots, W_J \right) = J^{-1} \sum_{b=1}^{J} (W_b - \overline{W})^2 = S_J^2.$$

Division by S_J therefore gives an exact conditional $N(0,1)$ variable. Neither distribution changes when an external within-block sample size increases while J remains fixed. □

Proof of Proposition C.2. The upper tail of a t_ν variable is regularly varying with index $-\nu$:

$$\mathbb{P}(t_\nu > x) \sim C_\nu x^{-\nu} \qquad (x \to \infty)$$

for a positive constant C_ν. Setting the tail probability equal to $\alpha/(2k)$ and inverting gives

$$t_{\nu, 1-\alpha/(2k)} \sim \left(\frac{2C_\nu k}{\alpha} \right)^{1/\nu},$$

which proves (C.1). The most stringent Holm step uses the same order of tail probability. For a standardized Gaussian vector $(Z_1, \ldots, Z_k)$, regardless of dependence,

$$\mathbb{P}\left(\max_{j \le k} |Z_j| > x \right) \le 2k e^{-x^2/2}.$$

Taking $x = \sqrt{2 \log(2k/\alpha)}$ proves the $O(\sqrt{\log k})$ upper order for a Gaussian critical value. □