UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 28

Beyond One-Step Distributed Maximum Likelihood: Local Bias Cancellation, Global Synchronization, and the Price of Inexact Curvature

A joint order–round frontier, its sharpness in an exact model, and measured communication costs

Simon-Pierre Boucher

with the AI co-author designation GPT-6.1 (prior exploratory draft; see the disclosure on page 3)

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

October 2026

Version 1.0

Abstract

Distributed maximum likelihood estimation trades statistical efficiency for communication. We study a two-phase protocol in which each of K workers corrects the bias of its local maximum likelihood estimator by an r-th order split-sample extrapolation over nested sub-blocks, the coordinator averages the corrected estimates, and q globally synchronized Newton corrections follow. For fixed-dimensional regular likelihoods with N observations we prove a joint order–round frontier: first-order equivalence to the centralized estimator holds whenever $K = o(N^{1-1/(2^{q+1}r)})$, so that one extra local extrapolation level and one extra synchronization round are exchangeable in the exponent, and the half-block jackknife followed by one Newton step tolerates $K = o(N^{7/8})$ partitions, against the $K = O(N^{1/2})$ regime stated for the plain-average one-step estimator. In the exponential model we derive exact bias, variance and Newton identities, show that the exponent is sharp for the construction with limiting shift $-\theta_0(\beta_r c^r)^{2^q}$ when $K \sim cN^{\alpha(r,q)}$, and characterize the Newton basin of attraction. A perturbation theorem for inexact curvature recovers the known single-machine-Hessian frontiers as special cases and gives the number of bits to which transmitted derivatives may be quantized at each local order. Monte Carlo experiments with paired replications (exponential, logistic and Poisson likelihoods), an exact-model boundary study, and timed multi-node benchmarks on an Apple-silicon cluster quantify the finite-sample variance cost of higher-order corrections and the wall-clock exchange rate between local fits and synchronization rounds. The components we combine are known; the joint frontier, its sharpness and the bit-precision result are new. No universal communication lower bound is claimed.

Keywords: distributed estimation, maximum likelihood, split-sample jackknife, one-step Newton estimator, communication complexity, asymptotic equivalence, Monte Carlo

JEL Classification: C13, C15, C55, C63

Authorship, provenance and AI-use disclosure

This working paper develops an exploratory research note dated 28 October 2026 whose listed authors were Simon-Pierre Boucher and "GPT-6.1", an AI co-author designation that the researcher asked to preserve; the note is frozen with checksums in the replication package (`references/initial_pamle/`). The present version was prepared by the researcher with Claude (Anthropic) acting as a research and programming assistant under his direction: the literature audit, proofs, code, experiments and text were generated in that collaboration and reviewed by the researcher, who bears sole scientific responsibility. The AI systems have no institutional affiliation, e-mail address or ORCID. Most journals do not accept AI systems as authors; for a journal submission the "GPT-6.1" designation should be moved from the author list to this disclosure, which is flagged here rather than silently changed. No funding line is claimed for this project.

1. Introduction

When N observations are stored on K machines, the centralized maximum likelihood estimator (MLE) $\widehat{\theta}_N$ is the natural benchmark: under regularity it attains the inverse Fisher information as asymptotic variance, but computing it requires either moving the data or iterating many synchronized passes over the network. The simplest distributed alternative fits the model on each block of $m = N/K$ observations and averages the K local estimates. Averaging reduces variance but not the $O(1/m)$ bias of a nonlinear local estimator, so the average is first-order equivalent to $\widehat{\theta}_N$ only when $K = o(N^{1/2})$ (Zhang et al., 2013; Battey et al., 2018; Rosenblatt and Nadler, 2016). Two ideas relax this barrier. One is *local*: debias each local estimate before averaging, by subsampling (Zhang et al., 2013), by an analytic bias formula (Gu and Chen, 2023; Yin, 2022) or by a jackknife (Wu et al., 2023). The other is *global*: broadcast the average and add one Newton step computed from aggregated scores and Hessians (Huang and Huo, 2019), or several rounds of gradient-only surrogate updates (Jordan et al., 2019; Fan et al., 2023). Each idea has its own cost: a local correction multiplies the number of local fits, a global correction adds a synchronization round and, for dense Hessians, an uplink of order p^2 numbers per worker.

This paper studies the two ideas jointly. Each worker computes the MLE on its block and on nested sub-blocks of sizes $m/s_2, \dots, m/s_r$, combines them with Richardson-type weights that cancel the first $r-1$ terms of the bias expansion while preserving the first-order influence function exactly, and uploads the corrected estimate. The coordinator averages, broadcasts the average and applies q Newton corrections from aggregated derivatives. Our question is quantitative: for fixed parameter dimension p, how fast may K grow with N while the resulting estimator remains first-order equivalent to $\widehat{\theta}_N$, how does the answer depend on (r, q), when is the answer sharp, and

what happens when the curvature used by the coordinator is inexact, stale or quantized?

Contributions.

1. *A joint order–round frontier* (Theorem 1). Under a uniform stochastic expansion with a bias expansion of order r and local Newton regularity, first-order equivalence to $\widehat{\theta}_N$ holds whenever

$$K = o\big(N^{\alpha(r,q)}\big), \qquad \alpha(r,q) := 1 - \frac{1}{2^{q+1}r}.$$

   The exponent depends on (r,q) only through the *effective depth* $D = 2^q r$: doubling the local order and adding one synchronization round are exchangeable. The half-block jackknife with one Newton step gives $K = o(N^{7/8})$; the plain average with one step gives $3/4$; the quarter-block extrapolation with one step gives $11/12$. The proof keeps the centered random remainder of the local corrections, of order $(m\sqrt{K})^{-1}$, which cannot be dropped even though its mean vanishes.
2. *Exact analysis and sharpness* (Theorem 2). In the exponential model we derive the exact bias and variance of the corrected local estimators from Gamma and Beta moments, an exact Newton error identity, the basin of attraction of the Newton map, and the boundary behaviour: if $K/N^{\alpha(r,q)} \to c \in (0,\infty)$ then $\sqrt{N}(\theta_r^{(q)} - \widehat{\theta}_N) \xrightarrow{p} -\theta_0(\beta_r c^r)^{2^q}$ for $q \geq 1$, where β_r is an explicit rational constant ($\beta_2 = -2$, giving the shift $-4\theta_0 c^4$ for the half-block jackknife with one step), and the estimator is not even $\sqrt{N}$-close to $\widehat{\theta}_N$ when $K/N^{\alpha(r,q)} \to \infty$. The sufficient exponent is therefore necessary for this construction.
3. *The price of inexact curvature* (Theorem 3). If the Newton step uses a Hessian within δ_N of $H_N(\widehat{\theta}_N)$ in operator norm, the admissible exponent changes in a computable way. Averaged local Hessians evaluated at the local optima (shipped once with the pilot) give $\alpha = 1 - 1/(2(r+q))$; a single machine's Hessian gives $\alpha = 1 - 1/(2r+q)$, which for $r = 1$ reproduces the $(t+1)/(t+2)$ rounds frontier of Fan et al. (2023) and the $2/3$ exponent of Jordan et al. (2019); relative quantization of the transmitted score and Hessian to b bits costs nothing asymptotically once $b \geq (\frac{1}{2} - r(1-\alpha))\log_2 N + \omega(1)$ when $K \sim N^{\alpha}$, so that higher local order permits coarser derivative uplinks.
4. *Evidence.* Paired-replication Monte Carlo experiments for exponential, logistic and Poisson likelihoods with N up to 2^{22}, an exact-model boundary study, ablations over split patterns, Newton rounds and derivative quantization, and timed benchmarks of the full coordinator–worker protocol on up to 16 Apple-silicon hosts quantify the finite-sample variance cost of higher-order corrections (which can reverse the asymptotic ranking at moderate m), the fraction of local fits that fail, interval coverage, transmitted bytes per round and the measured exchange rate between local fits and synchronization rounds.

What is not claimed. The two building blocks are known: split-sample extrapolation is the generalized jackknife of Quenouille (1956) and Schucany et al. (1971), applied to panels by Dhaene and Jochmans (2015); one-step distributed Newton estimation is due to Huang and Huo (2019). "Debias locally, then average" exists in several forms (Zhang et al., 2013; Gu and Chen, 2023; Wu et al., 2023; Yin, 2022), and "average, then correct globally" in several others (Huang and Huo, 2019; Zhu et al., 2021; Jordan et al., 2019; Fan et al., 2023). What we add is the joint analysis, the sharpness statement, and the inexact-curvature frontier. We claim no minimax or communication lower bound: in exponential families the pooled sufficient statistic recovers $\widehat{\theta}_N$ exactly with one round of p numbers per worker, and the exponential model is used here as an analytically transparent laboratory, not as an application. All results are for fixed p and i.i.d. data; constants depend on p and the theory does not transfer to growing dimension without new arguments. Wall-clock statements are measurements on specific hardware and are reported as such.

The paper is organized as follows. Section 2 positions the results against the closest theorems in the literature. Section 3 defines the likelihood environment, the protocol and the assumptions. Section 4 develops the general local correction. Section 5 states the frontier theorem and its inferential corollary, Section 6 the exact exponential analysis, Section 7 the inexact-curvature theorem, Section 8 the cost model. Sections 9 and 10 report the Monte Carlo and benchmark evidence, Section 11 the extensions, Section 12 the limitations. Proofs are in Sections A and B; replication material is described in Section D.

2. Related literature and positioning

Table 1 compares the closest results at the level of their theorems; the comparison was made on the full texts, with the caveats listed in the replication package when a published version could not be accessed.

One-shot averaging. Zhang et al. (2013) bound the mean squared error of the average by $O(1/N + 1/m^2)$ (their Theorem 1), which matches the $O(1/N)$ centralized rate only for $K \lesssim N^{1/2}$; Rosenblatt and Nadler (2016) show that the second-order bias of the average equals m times that of $\widehat{\theta}_N$ (their Theorem 2); Battey et al. (2018) obtain the same $\sqrt{N}$ barrier in sparse high dimension; Volgushev et al. (2019) prove for quantile regression that $K = o(N^{1/2})$ is also *necessary* for the average to be first-order equivalent, by exhibiting a model in which the remainder is exactly of order $1/m$ (their Theorem 3.2), and Shang and Cheng (2017) establish a sharp "maximal number of machines" for smoothing splines. Lin and Xi (2011) weight the local estimates by local Jacobians evaluated at the local estimates; their equivalence theorem requires $K \ll N^{1/4}$.

Local debiasing. The subsampled-average estimator of Zhang et al. (2013) combines each local fit with a fit on a random subsample of size $\lceil \rho m \rceil$ as $(\bar{\theta}_1 - \rho\bar{\theta}_2)/(1-\rho)$, a one-level Richardson extrapolation that removes the $1/m^2$ term of the MSE at the price of a leading constant $(2+3\rho)/(1-\rho)^2$ and a tuning rate $\rho \to 0$ (their Theorem 4); they leave the use of several subsamples open. Gu and Chen (2023) estimate the leading bias analytically on each machine with sample splitting and weight by local information, obtaining the centralized χ^2 limit for $K = o(m^2)$, i.e. $K = o(N^{2/3})$ (their Theorems 4–6), with a smaller constant than the subsampled average. Wu et al. (2023) apply the delete-one jackknife within each subsample and average, with bias $O(m^{-2}) + O(N^{-1})$ for smooth functions of means (their Theorem 2). The split-sample weights themselves go back to Quenouille (1949, 1956) and Schucany et al. (1971); Dhaene and Jochmans (2015) develop the general higher-order split-panel jackknife, with weights solving a Vandermonde-type system (their equations 5.1–5.3), prove that the first-order variance is unchanged and that partitions into $\{2,3,\dots,h+1\}$ parts minimize the inflation of the uncancelled bias terms, and leave the higher-order variance open. Hahn et al. (2024) show that in i.i.d. cross-sections the split-sample bias estimate is inconsistent and that split-sample corrections carry twice the higher-order variance of the delete-one jackknife (their Propositions 5–7), a cost that our exact variance formula and Monte Carlo results make explicit for nested splits.

Global corrections. Huang and Huo (2019) define the distributed one-step estimator: workers send scores and Hessians evaluated at the broadcast average and the coordinator takes one Newton step. Their Theorem 3.4 proves asymptotic equivalence to the full-sample M-estimator when $k = O(\sqrt{N})$; their Theorem 3.5 bounds the MSE by $2\operatorname{tr}\Sigma/N + O(N^{-2}) + O(k^4N^{-4})$, a bound that is itself $o(N^{-1})$-exact whenever $k = o(N^{3/4})$, although that exponent is not stated. Zhu et al. (2021) prove that one *local* Newton step per machine followed by information-weighted averaging is efficient iff $m/N^{1/4} \to \infty$, "allowing $K = O(N^{3/4})$" (their Theorem 2), and iterate the scheme. Jordan et al. (2019) replace the aggregated Hessian by that of the first machine and send gradients only; their Theorem 3.1 bounds the one-round error by $C(\|\bar{\theta} - \widehat{\theta}_N\|^2 + (N^{-1/2} + m^{-1/2})\|\bar{\theta} - \widehat{\theta}_N\|)$, a contraction by $m^{-1/2}$ per round. Fan et al. (2023) add a proximal term and state the explicit rounds frontier (their remark after Theorem 4.1): from the plain average, t single-machine-Hessian steps match the global minimizer when $m^{t+2} \gg N \cdot \text{polylog}$, i.e. $K = o(N^{(t+1)/(t+2)})$ up to logarithms, and one averaged local-Hessian step in the linear model allows $K = o(N^{3/4})$. Shamir et al. (2014) and Wang et al. (2017) are the optimization-side precursors; Chen et al. (2022) avoid the Hessian entirely by stochastic first-order iterations on one machine. Yin (2022) removes the quadratic terms of the one-step estimator analytically and obtains efficiency iff $K = o(N^{2/3})$. Reviews by Gao et al. (2022) and Li et al. (2024) confirm that no joint order–round frontier was available at their writing.

Table 1. Closest results, compared at theorem level (our notation: K blocks of size $m = N/K$).

Paper	Local correction	Global correction (curvature)	Uplink per w
Zhang et al. (2013) Thm 1	none	none	p
Zhang et al. (2013) Thm 4	one random subsample	none	p
Lin and Xi (2011) Thm 4.2	none	local-Jacobian weighting	$p+p^2$
Huang and Huo (2019) Thm 3.4/3.5	none	1 step, aggregated Hessian	p; $p+p^2$
Zhu et al. (2021) Thm 2	none	1 local step + weighting	$p+p^2$
Jordan et al. (2019) Cor 3.4	none	t rounds, machine-1 Hessian	p per round
Fan et al. (2023) rem. after Thm 4.1	none	t rounds, machine-1 Hessian	p per round
Gu and Chen (2023) Thm 4–6	analytic bias, split	information weighting	$p+p^2$
Wu et al. (2023) Thm 2–4	delete-one jackknife	none	p
Yin (2022) Thm 3.2	none	1 step + analytic adjustment	p + summarie
Volgushev et al. (2019) Cor 3.3	none	none	p
This paper, Theorem 1	order-r split extrapolation	q steps, aggregated Hessian	p; $q(p+p(p$
This paper, Theorem 3	order r	q steps, stale / machine-1 / b-bit Hessian	see Table 4

Regimes are sufficient conditions unless marked necessary; "MSE rate" means attainment of the $O(1/N)$ rate rather than $\sqrt{N}$-equiva
accepted manuscripts where publisher pages were inaccessible; see `literature/PRIOR_ART_MATRIX.csv`).

Gap. None of these papers combines an extrapolated local start with aggregated-Hessian Newton rounds, states a frontier indexed by both the local order and the number of rounds, proves sharpness for a corrected construction, or expresses the effect of inexact or quantized Hessians as a change in the admissible exponent with a bit count. These are the four results of Sections 5 to 7; the frontiers of Huang and Huo (2019), Zhu et al. (2021), Jordan et al. (2019) and Fan et al. (2023) appear as the cells $(1,1)$, $(1,1)$ with local Hessians, and $(1,t)$ with a single-machine Hessian of our tables.

3. Environment, protocol and assumptions

3.1. Likelihood, partitions and nested splits

Let $X_1,\dots,X_N$ be i.i.d. with density $f(\cdot;\theta_0)$ with respect to a dominating measure, θ_0 interior to an open $\Theta \subset \mathbb{R}^p$ and p fixed. For a sub-sample $S \subset \{1,\dots,N\}$ write $L_S(\theta) = |S|^{-1}\sum_{i\in S}\log f(X_i;\theta)$, $g_S = \nabla L_S$, $H_S = \nabla^2 L_S$ and let $\widehat{\theta}_S$ denote a measurable maximizer of L_S (the MLE on S). The centralized MLE is $\widehat{\theta}_N = \widehat{\theta}_{\{1,\dots,N\}}$. The index set is partitioned, independently of the data, into K blocks $D_1,\dots,D_K$ of equal size $m = N/K$; Section 11 treats unequal sizes. Let $1 = s_1 < s_2 < \cdots < s_r$ be integers, each dividing m. Level j of worker k consists of the s_j disjoint contiguous sub-blocks of D_k of size m/s_j; $T_{s_j,k}$ denotes the average of the s_j MLEs computed on them, so that $T_{1,k} = \widehat{\theta}_{D_k}$. The default pattern is $s = (1,2,\dots,2^{r-1})$; the pattern $(1,2,3)$ of Dhaene and Jochmans (2015) is also used. The collection of all sub-blocks at all levels, all blocks and the full sample is the *sub-sample family* $\mathscr{S}_N$. Asymptotics are along a triangular array $N \to \infty$ with $K = K_N$ and $m \to \infty$.

Algorithm 1 Order-r local correction with q global Newton corrections

Require: blocks $D_1, \dots, D_K$ on K workers; splits $s_1 < \cdots < s_r$ with weights w; rounds q

1: **Phase 1 (workers, in parallel).** For each level j, fit the MLE on each of the s_j sub-blocks; set $T_{s_j,k}$ to their average; upload $J_{r,k} = \sum_j w_j T_{s_j,k}$ (and a validity flag) ▷ p numbers up
2: **Coordinator.** $\theta \leftarrow K^{-1} \sum_k J_{r,k}$
3: **for** $t = 1, \dots, q$ **do** ▷ synchronized derivative phase
4: broadcast θ ▷ p numbers down to each worker
5: each worker uploads $g_{D_k}(\theta)$ and the upper triangle of $H_{D_k}(\theta)$ ▷ $p + p(p+1)/2$ numbers up
6: $g \leftarrow K^{-1} \sum_k g_{D_k}(\theta)$, $H \leftarrow K^{-1} \sum_k H_{D_k}(\theta)$; solve $H\Delta = g$; $\theta \leftarrow \theta - \Delta$
7: **end for**
8: **return** $\theta_r^{(q)} = \theta$, with the condition number of $-H$ and $\|g\|$ as diagnostics

3.2. The protocol

Algorithm 1 and Figure 1 describe the estimator. Phase 1 is local: worker k fits the $\sum_j s_j$ MLEs of its levels and uploads the p-vector

$$J_{r,k} = \sum_{j=1}^{r} w_j \, T_{s_j,k}, \tag{1}$$

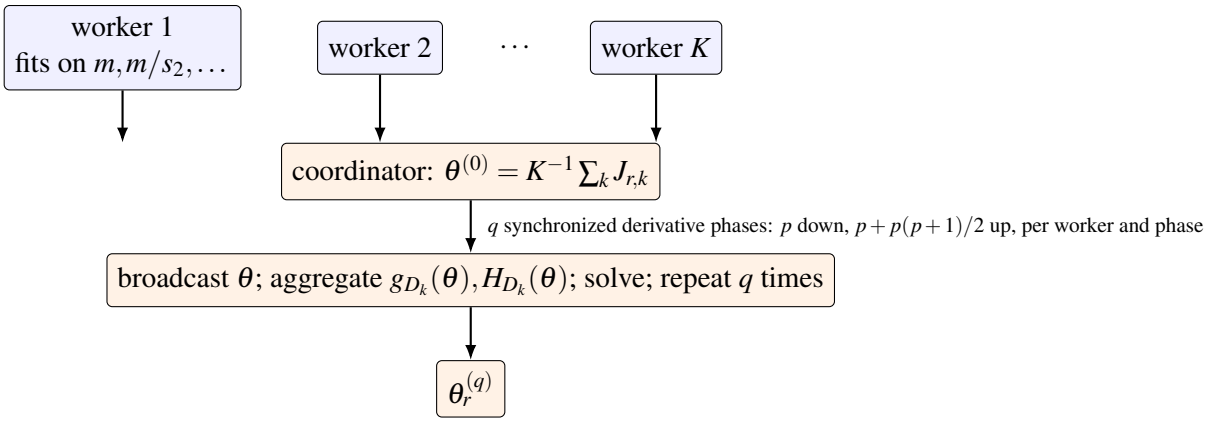
where $w = (w_1, \dots, w_r)$ solves the Vandermonde system of Section 4. The coordinator forms the pilot $\theta_r^{(0)} = K^{-1} \sum_k J_{r,k}$. Each of the q global corrections is one synchronized derivative phase: the coordinator broadcasts the current iterate θ, each worker uploads $g_{D_k}(\theta)$ and the $p(p+1)/2$ distinct entries of $H_{D_k}(\theta)$, the coordinator averages and solves

$$\theta \leftarrow \theta - H_N(\theta)^{-1} g_N(\theta), \tag{2}$$

by a linear solve, never by an explicit inverse. We write $\theta_r^{(q)}$ for the result; $(r,q) = (1,0)$ is the plain average, $(1,1)$ the one-step estimator of Huang and Huo (2019), $(2,1)$ the half-block jackknife with one step. Because the derivatives are evaluated at a common data-dependent point, the q derivative phases cannot be merged with phase 1: the protocol has $1+q$ synchronized exchange phases, $K(1+2q)$ messages, $K(p+q(p+p(p+1)/2))$ uplinked numbers and Kqp downlinked numbers, before headers and serialization (Table 4). Variants in which the coordinator uses a stale, single-machine or quantized Hessian are analysed in Section 7.

3.3. Assumptions

The theorems are conditional on the following assumptions, stated for the sub-sample family $\mathscr{S}_N$. They are high-level: they are implied by standard primitive conditions for regular parametric models (smoothness, moment and identification conditions of the type used by Zhang et al., 2013 and Huang and Huo, 2019, together with the validity of the formal bias expansion, Bhattacharya and



Phase 1 uplink: p numbers per worker (plus $p(p+1)/2$ if the local Hessian is shipped for the stale-curvature variant). Arrows are logical exchanges, not measured latency; Section 10 reports measured times and bytes.

Figure 1. Coordinator–worker protocol with exact uplink and downlink accounting.

Ghosh, 1978), and they are verified exactly in the exponential model of Section 6.

Assumption 1 (Uniform stochastic expansion)**.** There exist a measurable ψ with $\mathbb{E}\psi(X) = 0$ and $\Sigma_\psi := \operatorname{Var}\psi(X)$ finite and positive definite, an integer n_0 and a constant $C < \infty$ such that for every $S \in \mathscr{S}_N$ with $|S| \geq n_0$,

$$\widehat{\theta}_S - \theta_0 = |S|^{-1} \sum_{i \in S} \psi(X_i) + R_S, \qquad \mathbb{E}\|R_S\|^2 \leq C\,|S|^{-2}. \tag{3}$$

For a correctly specified regular likelihood $\psi = I(\theta_0)^{-1}\nabla_\theta \log f(X;\theta_0)$ and $\Sigma_\psi = I(\theta_0)^{-1}$; for a misspecified model θ_0 is the pseudo-true value, $\psi = -\bar{H}(\theta_0)^{-1}\nabla_\theta \log f(X;\theta_0)$ and Σ_ψ is the sandwich (Section 11). The same ψ serves every nested sub-sample; this is the only property of the influence function used.

Assumption 2 (Bias expansion of order r)**.** There exist vectors $b_1, \ldots, b_{r-1}$ and $C_b < \infty$ with

$$\mathbb{E}R_S = \sum_{a=1}^{r-1} b_a\,|S|^{-a} + \rho_{|S|}, \qquad \|\rho_n\| \leq C_b\,n^{-r}, \qquad S \in \mathscr{S}_N,\ |S| \geq n_0. \tag{4}$$

For $r = 1$ the assumption only asks $\|\mathbb{E}R_S\| \leq C_b|S|^{-1}$, which follows from **??** 1 by Jensen's inequality. The existence of the expansion for $r \geq 2$ is not implied by consistency or asymptotic normality; it requires smoothness of the likelihood to order $r+1$ and moment conditions, and it fails for estimators with non-smooth criteria.

Assumption 3 (Local Newton regularity)**.** With probability tending to one, $\widehat{\theta}_N$ is an interior zero of g_N; there are a fixed convex neighbourhood U of θ_0 and random variables $M_N, \Lambda_N = O_p(1)$

such that $\sup_{\theta\in U}\|H_N(\theta)^{-1}\| \le M_N$ and $\|H_N(u)-H_N(v)\| \le \Lambda_N\|u-v\|$ for $u,v\in U$. Moreover $\sqrt{N}(\widehat{\theta}_N-\theta_0) \Rightarrow \mathcal{N}(0,\Sigma_\psi)$.

Assumption 4 (Observation-level curvature smoothness; used only in Theorem 3(a))**.** $h(x;\theta) := \nabla^2_\theta \log f(x;\theta)$ is continuously differentiable on U with $\|\nabla h(x;\theta)-\nabla h(x;\theta')\| \le L(x)\|\theta-\theta'\|$, $\mathbb{E}L(X)^2<\infty$, $\mathbb{E}\|h(X;\theta_0)\|^2<\infty$, $\mathbb{E}\|\nabla h(X;\theta_0)\|^2<\infty$, and $\mathbb{E}\|\widehat{\theta}_{D_k}-\theta_0\|^4 \le C_4 m^{-2}$.

Remark 1. The assumptions fix p, require independent identically distributed blocks and well-behaved local optimizers. Constants such as C, M_N and Λ_N depend on p and on the parametrization; none of the exponents below should be read as valid when p grows with N. The condition $m\to\infty$ is essential: all statements are about a triangular array in which the number of partitions and the local sample size grow together, not about a fixed data set with a maximal number of machines.

4. The general local correction

4.1. Weights

Let V be the $r\times r$ Vandermonde matrix with entries $V_{aj}=s_j^a$, $a=0,\dots,r-1$. The weights are the unique solution of

$$\sum_{j=1}^{r} w_j = 1, \qquad \sum_{j=1}^{r} w_j s_j^a = 0, \quad a=1,\dots,r-1, \tag{5}$$

i.e. $w=V^{-1}e_1$. For $s=(1,2)$, $w=(2,-1)$; for $(1,2,4)$, $w=(8/3,-2,1/3)$; for $(1,2,4,8)$, $w=(64/21,-8/3,2/3,-1/21)$; for $(1,2,3)$, $w=(3,-3,1)$. Two constants summarize the cost of a pattern: the *leading bias coefficient* $\beta_r := \sum_j w_j s_j^r$, which multiplies the first uncancelled bias term, and the *remainder amplification* $A_r := \sum_j |w_j| s_j$, which multiplies the L_2 bound of the stochastic remainder. Table 2 lists them; the weights are computed exactly in rational arithmetic in the replication code and the moment constraints are verified to machine precision.

4.2. Influence preservation, bias cancellation and the random remainder

Write $U_S := |S|^{-1}\sum_{i\in S}\psi(X_i)$ and $U_k := U_{D_k}$. The sub-blocks of level j partition D_k exactly, so the average of their influence terms is U_k whatever j; since the weights sum to one, the influence term of $J_{r,k}$ is U_k (Lemma 1). Nothing else about ψ is used, which is why the construction applies verbatim to pseudo-MLEs. Collecting the remainders of the $\sum_j s_j$ fits into

$$R_k^J := \sum_{j=1}^{r} w_j s_j^{-1} \sum_{\ell=1}^{s_j} R_{S_{j\ell,k}}, \tag{6}$$

Table 2. Split patterns, weights and constants.

Pattern s	r	Weights w	β_r	A_r	$\sum_j \lvert w_j \rvert$	$\kappa(V)$	Local fits
(1)	1	1	1	1	1	1	1
$(1,2)$	2	$2, -1$	-2	4	3	6.9	3
$(1,3)$	2	$3/2, -1/2$	-3	3	2	5.8	4
$(1,2,4)$	3	$8/3, -2, 1/3$	8	8	5	80	7
$(1,2,3)$	3	$3, -3, 1$	6	12	7	71	6
$(1,2,4,8)$	4	$64/21, -8/3, 2/3, -1/21$	-64	$80/7$	6.4	3284	15
$(1,2,3,4)$	4	$4, -6, 4, -1$	-24	32	15	1171	10

$\kappa(V)$ is the spectral condition number of the Vandermonde matrix; "local fits" is $\sum_j s_j$ (each level costs one pass over the block's data). Source: `results/tables/weights_analysis.csv`.

we have $J_{r,k} - \theta_0 = U_k + R_k^J$. The mean of R_k^J is deterministic and, by (4) and (5), equals $\sum_{a<r} b_a m^{-a} \sum_j w_j s_j^a + \sum_j w_j \rho_{m/s_j} = O(m^{-r})$ (Lemma 2). Its L_2 norm is at most $\sqrt{C} A_r / m$ by the triangle inequality (Lemma 3); the nested fits of one worker are dependent, and no independence is used at this step. Across workers the blocks are independent, so the centered average has second moment at most $C A_r^2 / (K m^2)$ (Lemma 4). Finally the full-sample expansion $\widehat{\theta}_N - \theta_0 = U_N + R_N$ with $U_N = K^{-1} \sum_k U_k$ cancels the common influence term exactly:

$$\theta_r^{(0)} - \widehat{\theta}_N = \mathbb{E} R_k^J + K^{-1} \sum_k (R_k^J - \mathbb{E} R_k^J) - R_N = O_p\Big(m^{-r} + \frac{1}{m\sqrt{K}} + \frac{1}{N}\Big), \tag{7}$$

(Lemma 5). The middle term is the centered random part of the local corrections. Its mean is zero but its magnitude, $\sqrt{K}/N$, exceeds N^{-1} by the factor $\sqrt{K}$ and must be carried through every subsequent step; a proof that drops it is incomplete even where the final exponent happens to be unaffected.

4.3. Conditioning and the choice of pattern

Higher order costs conditioning. The weights grow geometrically with r for the doubling pattern ($\max_j |w_j| \approx 3.0, 3.2$ for $r = 4, 5$, with Vandermonde condition numbers 3×10^3 and 5×10^5), and the finest sub-block has $m/2^{r-1}$ observations, which for $p > 1$ quickly reaches the regime where local MLEs fail to exist (separation in logistic regression) or have no moments (sub-blocks of two or three exponential observations). The patterns of Dhaene and Jochmans (2015), $(1,2,3)$ and $(1,2,3,4)$, use coarser finest blocks and smaller $|\beta_r|$ but larger A_r (12 against 8 for $r = 3$): they trade bias constants against remainder amplification. Which trade is better is a finite-sample question answered in Section 9. Adaptive selection of r from the data is not covered by the theory: a data-driven order changes the estimator and its limit, and the theorems below are for fixed (r, q);

see Section 12.

5. Main results: the order–round frontier

5.1. One Newton step contracts quadratically

The global correction is a classical one-step estimator (Le Cam, 1956; Bickel, 1975; van der Vaart, 1998), with the difference that the comparison point is $\widehat{\theta}_N$ rather than θ_0. Under **??** 3, for any $\vartheta_N \xrightarrow{p} \theta_0$ the Newton map $\mathcal{N}_N(\vartheta) := \vartheta - H_N(\vartheta)^{-1} g_N(\vartheta)$ satisfies

$$\|\mathcal{N}_N(\vartheta_N) - \widehat{\theta}_N\| \le \tfrac{1}{2} M_N \Lambda_N \|\vartheta_N - \widehat{\theta}_N\|^2 \tag{8}$$

on an event of probability tending to one (Lemma 6), and q iterates satisfy $\|\theta^{(q)} - \widehat{\theta}_N\| \le (\tfrac{1}{2} M_N \Lambda_N)^{2^q-1} \|\theta^{(0)} - \widehat{\theta}_N\|^{2^q}$ (Lemma 7). Applied to the pilot bound (7), repeated squaring acts on all three terms:

$$\|\theta_r^{(q)} - \widehat{\theta}_N\| = O_p\Big(\Big[m^{-r} + \frac{1}{m\sqrt{K}} + \frac{1}{N} \Big]^{2^q} \Big). \tag{9}$$

5.2. The frontier

Theorem 1 (Order–round frontier)**.** *Let* **??** *1–3 hold with order $r \ge 1$ and fixed $q \ge 0$, let $m \to \infty$ and*

$$K = o\big(N^{\alpha(r,q)}\big), \qquad \alpha(r,q) = 1 - \frac{1}{2^{q+1} r}.$$

Then $\sqrt{N}\,(\theta_r^{(q)} - \widehat{\theta}_N) \xrightarrow{p} 0$, and consequently $\sqrt{N}(\theta_r^{(q)} - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_\psi)$.

Proof sketch. By (9), $\sqrt{N}\|\theta_r^{(q)} - \widehat{\theta}_N\|$ is bounded by a sum of three terms. The bias term is $\sqrt{N} m^{-r 2^q} = N^{1/2} (K/N)^{r 2^q}$, which vanishes iff $K^{r 2^q} = o(N^{r 2^q - 1/2})$, i.e. iff $K = o(N^{\alpha(r,q)})$. The random term is $\sqrt{N}(\sqrt{K}/N)^{2^q}$: for $q = 0$ it equals $\sqrt{K/N}$ and vanishes because $\alpha(r,0) < 1$; for $q \ge 1$ it is at most $N^{1/2 - 2^{q-1}} \le N^{-1/2}$ whenever $K \le N$. The full-sample term $\sqrt{N} N^{-2^q}$ vanishes. Consistency of the pilot, needed to apply (8), follows from (7) and $m \to \infty$. The limit law follows from **??** 3 and Slutsky's lemma. Section A gives the details, including the event on which all iterates stay in U. □

Table 3 displays the frontier. Three readings are useful. First, the cells $(1,0)$, $(1,1)$ and $(1,1)$-with-local-curvature are the regimes of Zhang et al. (2013), Huang and Huo (2019) (whose stated condition is the weaker $K = O(N^{1/2})$, while the $k^4 N^{-4}$ remainder of their MSE bound already implies $3/4$) and Zhu et al. (2021); the table extends them in both directions. Second, the exponent is symmetric in the exchange "one more round $\leftrightarrow$ double the local order": $(2,1)$, $(1,2)$ and $(4,0)$ all give $7/8$. Which of the three is cheaper is a question about the hardware, not about statistics,

Table 3. Sufficient exponents $\alpha(r,q)$ in $K = o(N^{\alpha})$ (Theorem 1), by local order r and number of aggregated Newton corrections q.

	$q=0$	$q=1$	$q=2$	$q=3$
$r=1$ (plain local MLE)	$1/2$	$3/4$	$7/8$	$15/16$
$r=2$ (half-block jackknife)	$3/4$	$7/8$	$15/16$	$31/32$
$r=3$ (quarter-block, or $(1,2,3)$)	$5/6$	$11/12$	$23/24$	$47/48$
$r=4$ (eighth-block, or $(1,2,3,4)$)	$7/8$	$15/16$	$31/32$	$63/64$

The exponent depends on (r,q) only through $D = 2^q r$; cells on an anti-diagonal of constant D coincide. These are sufficient conditions under fixed p; they are not machine counts for a given N, not minimax statements and not evidence of speed.

and is the subject of Sections 8 and 10. Third, the dominant restriction is always the deterministic bias term; the random remainder and the full-sample term are never binding for $q \geq 1$ and bind only through $K = o(N)$ for $q = 0$. This is a statement about rates, not about finite samples: in Section 9 the random term is visible as variance inflation at moderate m.

5.3. Inference

Corollary 1 (Inferential validity)**.** *Under the conditions of Theorem 1, let* $\widehat{\Sigma}_N \xrightarrow{p} \Sigma_\psi$. *Then for every* j *and level* $1-\tau$ *the interval* $\theta_r^{(q,j)} \pm z_{1-\tau/2}(\widehat{\Sigma}_{N,jj}/N)^{1/2}$ *has asymptotic coverage* $1-\tau$. *Under correct specification* $\widehat{\Sigma}_N = -H_N(\theta_r^{(q)})^{-1}$, *which is available at no extra communication from the last derivative phase (for* $q \geq 1$*); under misspecification the sandwich* $H_N^{-1}\widehat{J}_N H_N^{-1}$ *with* $\widehat{J}_N = K^{-1}\sum_k |D_k|^{-1}\sum_{i\in D_k} \nabla \log f(X_i;\theta)\nabla \log f(X_i;\theta)^\top$ *requires one additional* $p(p+1)/2$ *upload per worker in that phase.*

The corollary says nothing about estimators outside the frontier: when $K/N^{\alpha(r,q)}$ does not vanish, the point estimator has a non-centered limit (Theorem 2) and the interval under-covers, as the coverage tables of Section 9 show for the plain average and the one-step estimator at small m.

5.4. What the theorem does not say

The theorem is a sufficient statement under high-level assumptions; it gives neither the finite-sample behaviour (the constants C, C_b, M_N, Λ_N, A_r and β_r all matter at moderate m) nor any comparison between protocols at equal cost. The exponent cannot be improved *for this construction* (Theorem 2), but other constructions can do better: in exponential families the pooled sufficient statistic reaches $\widehat{\theta}_N$ exactly for every $K \leq N$, and multi-round schemes with $q \to \infty$ remove the restriction on K entirely at the price of unbounded synchronization (Jordan et al., 2019; Fan et al., 2023).

6. Exact analysis in the exponential model and sharpness

Let $X_i \sim \mathrm{Exp}(\theta_0)$ in the rate parametrization. The average log-likelihood is $L_n(\theta) = \log\theta - \theta\bar{X}_n$, so $\widehat{\theta}_n = 1/\bar{X}_n$, $g_n(\theta) = 1/\theta - \bar{X}_n$ and $H_n(\theta) = -1/\theta^2$. The model is trivial for distributed inference, since $\sum_{i\in D_k} X_i$ is a sufficient statistic that any worker can upload, and the Hessian does not depend on the data, so that single-machine curvature is exact. It is used here because every quantity in Sections 4 and 5 has a closed form, which makes it possible to check the assumptions, the constants and the boundary behaviour rather than assume them. The following statements are proved in Section B from the Gamma and Beta moments and verified symbolically in the replication package.

Theorem 2 (Exact exponential identities and boundary behaviour)**.** *Let* $n\bar{X}_n \sim \mathrm{Gamma}(n,\theta_0)$, $\psi(x) = \theta_0 - \theta_0^2 x$ *and* $R_n = \widehat{\theta}_n - \theta_0 - n^{-1}\sum_i \psi(X_i) = (1-\theta_0\bar{X}_n)^2/\bar{X}_n$.

(i) For $n > 2$*:* $\mathbb{E}\widehat{\theta}_n = \theta_0 n/(n-1)$, $\mathrm{Var}\,\widehat{\theta}_n = \theta_0^2 n^2/((n-1)^2(n-2))$, $\mathbb{E}R_n = \theta_0/(n-1) = \theta_0\sum_{a\geq 1} n^{-a}$ *and* $\mathbb{E}R_n^2 = \theta_0^2(3n+2)/(n(n-1)(n-2))$*, with* $\sup_{n\geq 3} n^2\mathbb{E}R_n^2/\theta_0^2 = 33/2$*. Hence* **??** *1***??** *2 hold for every* r *with* $n_0 = 3$, $C = 33\theta_0^2/2$ *and* $b_a = \theta_0$.

(ii) For any pattern with $m/s_r > 1$*:* $\mathbb{E}J_{r,k} - \theta_0 = \theta_0\sum_j w_j s_j/(m-s_j) = \beta_r\theta_0 m^{-r}(1+O(m^{-1}))$*; in particular* $\mathbb{E}J_{2,k} - \theta_0 = -2\theta_0/((m-1)(m-2))$ *and* $\mathbb{E}J_{3,k} - \theta_0 = 8\theta_0/((m-1)(m-2)(m-4))$.

(iii) For $m > 4$*:* $\mathrm{Var}\,J_{2,k} = \theta_0^2 m^2(m^2-6m+17)/((m-4)(m-2)^2(m-1)^2) = \theta_0^2(m^{-1}+4m^{-2}+O(m^{-3}))$*, the same first two terms as* $\mathrm{Var}\,\widehat{\theta}_m$*; the variance cost of the half-block correction is* $O(m^{-3})$ *in this model.*

(iv) Newton identities: $\mathcal{N}_N(x) = 2x - x^2\bar{X}_N$ *and* $\mathcal{N}_N(x) - \widehat{\theta}_N = -(x-\widehat{\theta}_N)^2/\widehat{\theta}_N$ *exactly; after* q *steps* $\mathcal{N}_N^q(x) - \widehat{\theta}_N = -(x-\widehat{\theta}_N)^{2^q}/\widehat{\theta}_N^{2^q-1}$*. The iteration converges iff* $0 < x < 2\widehat{\theta}_N$ *and diverges to* $-\infty$ *otherwise (*$x = 2\widehat{\theta}_N$ *maps to* 0 *and stays there).*

(v) Boundary: if $m\to\infty$ *and* $K/N^{\alpha(r,q)} \to c \in (0,\infty)$*, then* $\sqrt{N}(\theta_r^{(q)} - \widehat{\theta}_N) \xrightarrow{p} -\theta_0(\beta_r c^r)^{2^q}$ *for* $q \geq 1$ *and* $\xrightarrow{p} \beta_r\theta_0 c^r$ *for* $q = 0$*; for* $(r,q) = (2,1)$ *the limit is* $-4\theta_0 c^4$*, so that* $\sqrt{N}(\theta_2^{(1)} - \theta_0) \Rightarrow \mathcal{N}(-4\theta_0 c^4, \theta_0^2)$.

(vi) Necessity: if $m\to\infty$ *and* $K/N^{\alpha(r,q)} \to \infty$*, then* $\sqrt{N}\,|\theta_r^{(q)} - \widehat{\theta}_N| \xrightarrow{p} \infty$.

Items (v)–(vi) show that the exponent of Theorem 1 is sharp for the construction in this model: the small-o cannot be weakened to O, and beyond the frontier the estimator is not $\sqrt{N}$-consistent for $\widehat{\theta}_N$. Three qualifications matter. The boundary regime $m \asymp N^{1/(2^{q+1}r)}$ is extremely slow to reach—for $(2,1)$, $m \asymp N^{1/8}$, so that $m \leq 16$ for all $N \leq 2^{24}$—and the finite-m bias $-2\theta_0/((m-1)(m-2))$ exceeds its asymptotic form $-2\theta_0 m^{-2}$ by a factor $m^2/((m-1)(m-2))$ that is 1.5 at $m = 6$; Section 9 therefore reports both the limit and the exact finite-m prediction. The sharpness is for this estimator family: it is not a lower bound over all distributed procedures, as pooling $\sum_i X_i$ shows. And the signs alternate: $\beta_2 = -2 < 0 < \beta_3 = 8$, so the higher-order pilot over-corrects where the half-block pilot under-corrects, a reason why finite-sample accuracy is not monotone in r.

Item (iv) also gives an explicit failure gallery. Because the weights have negative entries, $J_{r,k}$ can be negative when a sub-block average is small; the pilot then lies outside the basin $(0, 2\widehat{\theta}_N)$ with positive probability for every finite m, and the Newton iteration diverges rather than converges. The probability is negligible for $m \geq 32$ and visible for $m \leq 8$ in Section 9, where such replications are reported, not discarded. A production implementation would damp or project the step; the theory above describes the undamped map.

7. The price of inexact curvature

The exact protocol uploads $p + p(p+1)/2$ numbers per worker in every derivative phase. Three cheaper variants are common in practice: a *stale* Hessian, equal to the average of the local Hessians evaluated at the local MLEs and shipped once together with the pilot, after which every round costs p numbers; a *single-machine* Hessian evaluated at the current iterate on one worker, as in the surrogate-likelihood approach of Jordan et al. (2019); and a *quantized* exact Hessian. The following perturbation theorem covers all three and inexact scores as well. Let ϑ be the current iterate, $e := \vartheta - \widehat{\theta}_N$, and let the coordinator compute $\tilde{\vartheta} := \vartheta - \widehat{H}^{-1}\widehat{g}$ with $\widehat{g} = g_N(\vartheta) + \delta_g$.

Theorem 3 (Inexact Newton steps)**.** *Let* **??** *1–3 hold,* $\vartheta \xrightarrow{p} \theta_0$, $\|\widehat{H} - H_N(\widehat{\theta}_N)\| = O_p(\delta_N)$ *with* $\delta_N \to 0$ *and* $\|\widehat{H}^{-1}\| = O_p(1)$. *Then*

$$\|\tilde{\vartheta} - \widehat{\theta}_N\| = O_p\big(\|e\|^2 + \delta_N\|e\| + \|\delta_g\|\big). \tag{10}$$

With q such steps from the order-r pilot and a Hessian error δ_N at every step, $\|\theta_r^{(q)} - \widehat{\theta}_N\| = O_p(\|e_0\|\delta_N^q + \|e_0\|^{2^q} + \text{cross terms})$, and $\sqrt{N}(\theta_r^{(q)} - \widehat{\theta}_N) \xrightarrow{p} 0$ provided $\sqrt{N}\, m^{-r}\delta_N^q \to 0$ together with the conditions of Theorem 1. In particular:

(a) (stale averaged local curvature) *under* **??** *4, $\widehat{H} = K^{-1}\sum_k H_{D_k}(\widehat{\theta}_{D_k})$ has $\delta_N = m^{-1} + N^{-1/2}$, and the sufficient exponent becomes $\alpha = 1 - 1/(2(r+q))$ for $q \geq 1$;*

(b) (single-machine curvature) *$\widehat{H} = H_{D_1}(\vartheta)$ has $\delta_N = m^{-1/2}$ (no third-derivative condition needed), and the sufficient exponent becomes $\alpha = 1 - 1/(2r+q)$;*

(c) (relative quantization) *if the score and the Hessian are transmitted with relative precision ε (every entry rounded to b significant bits, $\varepsilon = 2^{-b}$), then (10) holds with $\delta_N = \varepsilon$ and $\|\delta_g\| = O_p(\varepsilon\|e\|)$, so that one quantized step from the order-r pilot is first-order equivalent to $\widehat{\theta}_N$ whenever $\sqrt{N}\,\varepsilon\, m^{-r} \to 0$ in addition to $K = o(N^{\alpha(r,1)})$. When $K \sim N^\alpha$ with $1 - 1/(2r) \leq \alpha < \alpha(r,1)$, it suffices that*

$$b \;\geq\; \Big(\tfrac{1}{2} - r(1-\alpha)\Big)\log_2 N + \omega(1), \tag{11}$$

and for $\alpha < 1 - 1/(2r)$ any fixed number of bits suffices (the pilot alone is already first-order

Table 4. Protocol variants: local fits, synchronized phases, uplink per worker, and sufficient exponents.

Variant	Curvature used by the coordinator	Local fits	Uplink per worker (numbers)
$(r,0)$	none	$\sum_j s_j$	p
(r,q) exact	aggregated $H_N(\theta)$ at each iterate	$\sum_j s_j$	$p+q\,(p+p(p+1)/2)$
(r,q) stale	$K^{-1}\sum_k H_{D_k}(\widehat{\theta}_{D_k})$, shipped once	$\sum_j s_j$	$p+p(p+1)/2+q\,p$
(r,q) single	$H_{D_1}(\theta)$ on worker 1	$\sum_j s_j$	$p+q\,p$ (+ none for curvature)
$(r,1)$ quantized	aggregated, b mantissa bits	$\sum_j s_j$	$p+(p+p(p+1)/2)\,b/64$ (in 64-bi
AEE (Lin and Xi, 2011)	local Jacobians at local estimates	1	$p+p(p+1)/2$
CSL (Jordan et al., 2019), t rounds	machine 1, full surrogate	1	$t\,p$
exact pooling (exponential families)	sufficient statistic	1	$\dim(T)$

Downlink per worker: qp numbers for every variant with $q \geq 1$. Headers, serialization and the pilot's validity flag are excluded; n conditions from Theorems 1 and 3; the CSL and AEE rows are the cited authors' results translated to our notation.

equivalent).

The proof (Section A) is a matrix-perturbation argument around the integral form of the Newton remainder: $\tilde{\vartheta} - \widehat{\theta}_N = \widehat{H}^{-1}[\widehat{H} - \int_0^1 H_N(\widehat{\theta}_N + te)\,\mathrm{d}t]\,e - \widehat{H}^{-1}\delta_g$, and the bracket is bounded by $\|\widehat{H} - H_N(\widehat{\theta}_N)\| + \frac{1}{2}\Lambda_N\|e\|$. The theorem turns the qualitative statements in the literature—the $m^{-1/2}$ factor of Jordan et al. (2019, Theorem 3.1), the $(\beta/\lambda)^2 \asymp m^{-1}$ contraction of averaged local Hessians in Shamir et al. (2014, Theorem 2) and Fan et al. (2023)—into exponents. Table 4 tabulates them with the communication they require. Case (b) with $r = 1$ gives $\alpha = 1 - 1/(2+q) = (q+1)/(q+2)$, exactly the rounds frontier that Fan et al. (2023) state for CEASE up to logarithms, and $q = 1$ gives the $2/3$ of Jordan et al. (2019); case (a) with $(r,q) = (1,1)$ gives $3/4$, the regime of Zhu et al. (2021). The stale variant has effective depth $r+q$ (linear in q) against $2^q r$ for exact curvature: with stale curvature a second round is worth one more local level, with exact curvature it is worth doubling the local order.

Case (c) is, to our knowledge, new in this form. It says that the precision needed for the derivative uplink is governed by how far the pilot already is from $\widehat{\theta}_N$: a pilot of order r with $K \sim N^\alpha$ has error $m^{-r} = N^{-r(1-\alpha)}$, and a relative error ε in the step costs $\varepsilon N^{-r(1-\alpha)}$, which is negligible at the $N^{-1/2}$ scale iff (11). For the half-block jackknife at $\alpha = 0.8$ the requirement is $0.1\log_2 N$ bits, i.e. two bits of mantissa at $N = 2^{20}$, whereas the plain average at the same α needs $0.3\log_2 N = 6$ bits and is outside its own frontier $(3/4)$ anyway. The mantissa bits multiply the $p(p+1)/2$ Hessian entries per worker and round, so the saving in bytes is $(64-b)/64$ of the Hessian payload. The statement is asymptotic in the same sense as Theorem 1, and the Monte Carlo ablation of Section 9 shows at which b the quantization error becomes visible at finite N.

8. The economics of synchronization

The frontier of Table 3 makes $(2,1)$, $(1,2)$ and $(4,0)$ statistically interchangeable at the level of exponents. Choosing among them is a question of cost, and the cost model below is proposed, calibrated and tested, not proved. Let $T_{\mathrm{fit}}(n,p)$ be the elapsed time of one local MLE fit on n observations, $T_{\mathrm{der}}(m,p)$ that of one local score-and-Hessian evaluation, λ the latency of one synchronized exchange (broadcast plus the slowest worker's reply, excluding payload), B the effective uplink bandwidth at the coordinator, b the bytes per transmitted number and $T_{\mathrm{solve}}(p)$ the coordinator's linear solve. For the doubling pattern the local phase of order r performs 2^r-1 fits touching every local observation r times, so that with a fit cost $T_{\mathrm{fit}}(n,p)\approx c_0(p)+c_1(p)\,n$ (fixed overhead plus a linear pass, the empirical shape of Newton fits with a bounded number of iterations),

$$T(r,q)\approx \underbrace{(2^r-1)c_0+r\,c_1 m}_{\text{local, critical path of the slowest worker}} + q\Big[\lambda+T_{\mathrm{der}}(m,p)+\frac{K\,b\,(p+p(p+1)/2)}{B}+T_{\mathrm{solve}}(p)\Big] + T_{\mathrm{sched}}(K)+T_{\mathrm{st}} \tag{12}$$

where the last two terms collect worker start-up and the excess of the slowest worker over the mean. Sums are critical-path sums: the local fits of different workers overlap, the q derivative phases do not.

Proposition 1 (Exchange rate between a local level and a round, under (12))**.** *Hold the effective depth $D=2^q r$ fixed. Moving from (r,q) to $(2r,q-1)$ removes one derivative phase and adds the local levels $r+1,\dots,2r$. Under (12) the round is cheaper iff*

$$\lambda+T_{\mathrm{der}}(m,p)+\frac{Kb(p+p(p+1)/2)}{B}+T_{\mathrm{solve}}(p)\;<\;(2^{2r}-2^r)\,c_0+r\,c_1 m. \tag{13}$$

The right-hand side grows linearly in m and exponentially in r; the left-hand side grows linearly in K and quadratically in p. For fixed N and small p the rule favours local levels when K is large and m is small, and favours rounds when m is large. Holding statistical accuracy fixed in finite samples requires, in addition, the variance inflation of higher local orders (Section 9) to be accounted for.

The proposition is a tautology given (12); its content is the calibration. Section 10 measures every term of (12) on the cluster—fit and derivative times by m and p, per-phase latency percentiles, bytes per phase, start-up and straggler excess—and reports where (13) lands for the designs run. We do not model overlapping of computation and communication, asynchronous updates (which change the estimator) or energy; these are listed as open in Section 12.

9. Monte Carlo evidence

Results are being generated; see Section D.

10. Measured distributed benchmarks

Benchmarks are being generated; see Section D.

11. Extensions: unequal blocks and misspecification

Both extensions are known in the literature (Zhu et al., 2021; Gu and Chen, 2023; Jordan et al., 2019) and are stated here only so that the frontier can be used in practice.

Proposition 2 (Unequal block sizes and pseudo-likelihood). *(i) Let the blocks have sizes $m_1, \dots, m_K$ with $\sum_k m_k = N$, $m_{\min} \to \infty$ and each $s_j \mid m_k$. With the size weights m_k/N in the pilot, $\theta_r^{(0)} - \widehat{\theta}_N = O_p(m_{\text{eff}}^{-r} + \sqrt{K}/N + N^{-1})$ with $m_{\text{eff}} := (N^{-1}\sum_k m_k^{1-r})^{-1/r} \geq m_{\min}$, and Theorem 1 holds with m replaced by m_{eff} (the aggregated derivatives are size-weighted and equal the full-sample derivatives, so the Newton steps are unchanged). With equal weights $1/K$ the influence term is not preserved: $K^{-1}\sum_k U_k - U_N = \sum_i c_i \psi(X_i)$ with $c_i = 1/(K m_{k(i)}) - 1/N$, a centered term of order $(\sum_i c_i^2)^{1/2}$ that is of exact order $N^{-1/2}$ unless the sizes are asymptotically equal; the equal-weight pilot is therefore not first-order equivalent for $q = 0$, while for $q \geq 1$ the discrepancy is of order $N^{-1/2}$ and is absorbed by the first Newton step.*

*(ii) If the model is misspecified and θ_0 is the pseudo-true value, all statements hold with $\psi = -\bar{H}(\theta_0)^{-1}\nabla \log f(X;\theta_0)$ and $\Sigma_\psi = \bar{H}(\theta_0)^{-1}\bar{J}(\theta_0)\bar{H}(\theta_0)^{-1}$, provided **??** 1–3 are read for the pseudo-MLE. The estimator is then first-order equivalent to the centralized pseudo-MLE and Wald intervals require the sandwich of Corollary 1; the observed-information intervals are not valid.*

The proof is in Section A; the Monte Carlo evidence is in Section 9 (skewed partitions with size ratios 4 and 16, and an omitted-nonlinearity logistic design). Heterogeneous distributions across blocks, in which the local pseudo-true values differ, are outside the scope: the common-influence identity of Lemma 1 fails, and the appropriate target and weighting are those of Gu and Chen (2023) and Duan et al. (2022).

12. Discussion and limitations

What is known and what is new. The local correction is the generalized jackknife, the global correction is the distributed one-step estimator, and the idea of debiasing before averaging has

analytic, delete-one and subsampling versions in the literature (Section 2). The joint frontier of Theorem 1, its sharpness in Theorem 2 and the inexact-curvature frontier with the bit requirement of Theorem 3 are the new statements; the Monte Carlo and the measured benchmarks are new evidence. We have found no prior statement of the exponents $7/8$ or $11/12$, of a frontier indexed by both r and q, or of a sharpness result for a bias-corrected construction, but we have not been able to check every unpublished thesis (the replication package lists the sources that could not be read in full).

Fixed dimension. All exponents are for fixed p. The constants C, C_b, M_N, Λ_N in **??** 1–3 depend on p, the Hessian payload grows as p^2, and the number of observations in the finest sub-block is m/s_r, so that for p comparable to m/s_r the local MLEs fail to exist with non-negligible probability (Section 9, $p = 20$). The results do not transfer to $p_N \to \infty$ without a new analysis of how every constant scales, and we make no claim in that direction.

Finite samples versus exponents. The frontier orders estimators by the exponent of the bias term alone. At moderate m the centered random remainder, whose constant grows with A_r, and the alternating signs of β_r make the order-3 pilot worse than the order-2 pilot in RMSE for all but the smallest m in our designs, and the one-step versions inherit part of this ranking. Doubling the local order to save a round is therefore not free statistically, in addition to its computational price; the higher-order variance results of Hahn et al. (2024) describe the same phenomenon for split-sample corrections of cross-section MLEs.

Adaptive order. Choosing r from the data (for instance by comparing successive pilots) changes the estimator, and nothing in Theorem 1 covers a data-dependent r. Sample splitting or a selection event with uniform bounds would be needed to restore a theorem; we leave this open and do not report an adaptive estimator.

Asynchrony, failures and stragglers. The theorems describe synchronized phases: every worker's derivatives are evaluated at the same broadcast point. Asynchronous or stale-iterate updates change the estimator and inherit no guarantee from the present analysis; worker failures were handled in the benchmarks by waiting, not by dropping workers, and the theory for dropped blocks is that of Huang and Huo (2019, Corollary 3.7) rather than ours.

Lower bounds. Nothing here is a minimax or communication lower bound. In exponential families the pooled sufficient statistic recovers $\widehat{\theta}_N$ exactly in one round of $\dim(T)$ numbers per worker, and the exponential model shows that the full surrogate-likelihood maximization is also

exact there. The sharpness in Theorem 2 concerns the estimator family $\{\theta_r^{(q)}\}$, not all distributed procedures.

Benchmarks. Measured times are specific to one cluster, one implementation (Python/NumPy over TCP with pickle framing) and the designs run. They calibrate (12) and show which side of (13) the designs fall on; they do not establish a general speed advantage of any protocol, and we report no energy or monetary cost because none was measured.

13. Conclusion

For regular fixed-dimensional likelihoods, local split-sample extrapolation of order r followed by q aggregated Newton corrections is first-order equivalent to the centralized MLE whenever $K = o(N^{1-1/(2^{q+1}r)})$, and this exponent is sharp for the construction in an exactly solvable model, where the limiting shift at the boundary is $-\theta_0(\beta_r c^r)^{2^q}$. One synchronization round and a doubling of the local order are exchangeable in the exponent, which turns the choice between them into a question about hardware that a calibrated cost model and measured benchmarks can answer for a given installation. Inexact curvature changes the exponent in a computable way, recovering the known single-machine-Hessian frontiers and showing that higher local order tolerates coarser quantization of the transmitted derivatives. The finite-sample evidence tempers the exponents: higher-order corrections inflate variance and can fail to exist at small block sizes, and the boundary regime is reached only at very large N. The components combined here are known; the joint frontier, its sharpness and the bit-precision result are the contribution.

References

Battey, Heather, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu, "Distributed Testing and Estimation under Sparse High Dimensional Models," *Annals of Statistics*, 2018, *46* (3), 1352–1382.

Bhattacharya, R. N. and J. K. Ghosh, "On the Validity of the Formal Edgeworth Expansion," *Annals of Statistics*, 1978, *6* (2), 434–451.

Bickel, P. J., "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association*, 1975, *70* (350), 428–434.

Cam, Lucien Le, "On the Asymptotic Theory of Estimation and Testing Hypotheses," in "Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability," Vol. 1, Berkeley: University of California Press, 1956, pp. 129–156.

Chen, Xi, Weidong Liu, and Yichen Zhang, "First-Order Newton-Type Estimator for Distributed Estimation and Inference," *Journal of the American Statistical Association*, 2022, *117* (540), 1858–1874.

Cordeiro, Gauss M. and Peter McCullagh, "Bias Correction in Generalized Linear Models," *Journal of the Royal Statistical Society: Series B*, 1991, *53* (3), 629–643.

Cox, D. R. and E. J. Snell, "A General Definition of Residuals," *Journal of the Royal Statistical Society: Series B*, 1968, *30* (2), 248–275.

Dhaene, Geert and Koen Jochmans, "Split-Panel Jackknife Estimation of Fixed-Effect Models," *Review of Economic Studies*, 2015, *82* (3), 991–1030.

Duan, Rui, Yang Ning, and Yong Chen, "Heterogeneity-Aware and Communication-Efficient Distributed Statistical Inference," *Biometrika*, 2022, *109* (1), 67–83.

Fan, Jianqing, Yongyi Guo, and Kaizheng Wang, "Communication-Efficient Accurate Statistical Estimation," *Journal of the American Statistical Association*, 2023, *118* (542), 1000–1010.

Gao, Yuan, Weidong Liu, Hansheng Wang, Xiaozhou Wang, Yibo Yan, and Riquan Zhang, "A Review of Distributed Statistical Inference," *Statistical Theory and Related Fields*, 2022, *6* (2), 89–99.

Gu, Jia and Songxi Chen, "Distributed Statistical Inference under Heterogeneity," *Journal of Machine Learning Research*, 2023, *24* (387), 1–56.

Hahn, Jinyong, David W. Hughes, Guido Kuersteiner, and Whitney K. Newey, "Efficient Bias Correction for Cross-Section and Panel Data," *Quantitative Economics*, 2024, *15* (3), 783–816.

Huang, Cheng and Xiaoming Huo, "A Distributed One-Step Estimator," *Mathematical Programming*, 2019, *174* (1–2), 41–76.

Jordan, Michael I., Jason D. Lee, and Yun Yang, "Communication-Efficient Distributed Statistical Inference," *Journal of the American Statistical Association*, 2019, *114* (526), 668–681.

Li, Xuetong, Yuan Gao, Hong Chang, Danyang Huang, Yingying Ma, Rui Pan, Haobo Qi, Feifei Wang, Shuyuan Wu, Ke Xu, Jing Zhou, Xuening Zhu, Yingqiu Zhu, and Hansheng Wang, "A Selective Review on Statistical Methods for Massive Data Computation: Distributed Computing, Subsampling, and Minibatch Techniques," 2024. arXiv:2403.11163.

Lin, Nan and Ruibin Xi, "Aggregated Estimating Equation Estimation," *Statistics and Its Interface*, 2011, *4* (1), 73–83.

Quenouille, M. H., "Approximate Tests of Correlation in Time-Series," *Journal of the Royal Statistical Society: Series B*, 1949, *11* (1), 68–84.

_ , "Notes on Bias in Estimation," *Biometrika*, 1956, *43* (3/4), 353–360.

Rosenblatt, Jonathan D. and Boaz Nadler, "On the Optimality of Averaging in Distributed Statistical Learning," *Information and Inference: A Journal of the IMA*, 2016, *5* (4), 379–404.

Schucany, W. R., H. L. Gray, and D. B. Owen, "On Bias Reduction in Estimation," *Journal of the American Statistical Association*, 1971, *66* (335), 524–533.

Shamir, Ohad, Nathan Srebro, and Tong Zhang, "Communication-Efficient Distributed Optimization Using an Approximate Newton-Type Method," in "Proceedings of the 31st International Conference on Machine Learning," Vol. 32 of *Proceedings of Machine Learning Research* PMLR 2014, pp. 1000–1008.

Shang, Zuofeng and Guang Cheng, "Computational Limits of a Distributed Algorithm for Smoothing Spline," *Journal of Machine Learning Research*, 2017, *18* (108), 1–37.

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

Volgushev, Stanislav, Shih-Kang Chao, and Guang Cheng, "Distributed Inference for Quantile Regression Processes," *Annals of Statistics*, 2019, *47* (3), 1634–1662.

Wang, Jialei, Mladen Kolar, Nathan Srebro, and Tong Zhang, "Efficient Distributed Learning with Sparsity," in "Proceedings of the 34th International Conference on Machine Learning," Vol. 70 of *Proceedings of Machine Learning Research* PMLR 2017, pp. 3636–3645.

Wu, Shuyuan, Xuening Zhu, and Hansheng Wang, "Subsampling and Jackknifing: A Practically Convenient Solution for Large Data Analysis with Limited Computational Resources," *Statistica Sinica*, 2023, *33*, 2041–2064.

Yin, Zhuoran, "Communication-Efficient Distributed Newton-Like Optimization with Gradients and M-Estimators," 2022. arXiv:2207.06253.

Zhang, Yuchen, John C. Duchi, and Martin J. Wainwright, "Communication-Efficient Algorithms for Statistical Optimization," *Journal of Machine Learning Research*, 2013, *14*, 3321–3363.

Zhu, Xuening, Feng Li, and Hansheng Wang, "Least-Square Approximation for a Distributed System," *Journal of Computational and Graphical Statistics*, 2021, *30* (4), 1004–1018.

Appendices

A. Proofs of the general results

Throughout, $S_{j\ell,k}$ ($\ell = 1,\dots,s_j$) denotes the ℓ-th sub-block of level j of worker k, of size m/s_j; $U_S = |S|^{-1}\sum_{i\in S}\psi(X_i)$; w solves (5).

Lemma 1 (Influence preservation)**.** *For every k and j, $s_j^{-1}\sum_{\ell=1}^{s_j} U_{S_{j\ell,k}} = U_k$, and hence $\sum_j w_j s_j^{-1}\sum_\ell U_{S_{j\ell,k}} = U_k$.*

Proof. The sub-blocks of level j partition D_k into s_j sets of equal size m/s_j, so $s_j^{-1}\sum_\ell U_{S_{j\ell,k}} = s_j^{-1}\sum_\ell (s_j/m)\sum_{i\in S_{j\ell,k}}\psi(X_i) = m^{-1}\sum_{i\in D_k}\psi(X_i) = U_k$. The second claim uses $\sum_j w_j = 1$. □

Lemma 2 (Bias cancellation)**.** *Under* **??** *1***??** *2 with $m/s_r \geq n_0$, $J_{r,k} - \theta_0 = U_k + R_k^J$ with R_k^J as in* (6)*, and*

$$\mathbb{E}R_k^J = \sum_{j=1}^{r} w_j \rho_{m/s_j}, \qquad \|\mathbb{E}R_k^J\| \leq C_b\, m^{-r}\sum_{j=1}^{r}|w_j|\, s_j^r.$$

In particular $\mathbb{E}R_k^J$ is deterministic and does not depend on k.

Proof. Applying (3) to each sub-block and Lemma 1, $J_{r,k} - \theta_0 = \sum_j w_j s_j^{-1} \sum_\ell (U_{S_{j\ell,k}} + R_{S_{j\ell,k}}) = U_k + R_k^J$. Taking expectations with (4) at $n = m/s_j$,

$$\mathbb{E}R_k^J = \sum_j w_j \Big[\sum_{a=1}^{r-1} b_a (s_j/m)^a + \rho_{m/s_j} \Big] = \sum_{a=1}^{r-1} b_a m^{-a} \sum_j w_j s_j^a + \sum_j w_j \rho_{m/s_j},$$

and the inner sums vanish by (5). The bound follows from $\|\rho_{m/s_j}\| \le C_b (s_j/m)^r$. □

Lemma 3 (Local remainder)**.** *Under* **??** *1,* $(\mathbb{E}\|R_k^J\|^2)^{1/2} \le \sqrt{C} A_r/m$ *with* $A_r = \sum_j |w_j| s_j$.

Proof. By Minkowski's inequality in L_2, $(\mathbb{E}\|R_k^J\|^2)^{1/2} \le \sum_j |w_j| s_j^{-1} \sum_\ell (\mathbb{E}\|R_{S_{j\ell,k}}\|^2)^{1/2} \le \sum_j |w_j| s_j^{-1} \cdot s_j \cdot \sqrt{C}(s_j/m) = \sqrt{C} A_r/m$. No independence between the nested sub-samples of a worker is used. □

Lemma 4 (Averaged centered remainder)**.** *Under* **??** *1,* $\mathbb{E}\big\|K^{-1} \sum_{k=1}^K (R_k^J - \mathbb{E}R_k^J)\big\|^2 \le C A_r^2/(Km^2)$.

Proof. The blocks are disjoint, so the vectors $R_k^J - \mathbb{E}R_k^J$ are independent and centered; the second moment of the average is $K^{-2} \sum_k \mathbb{E}\|R_k^J - \mathbb{E}R_k^J\|^2 \le K^{-1} \sup_k \mathbb{E}\|R_k^J\|^2 \le C A_r^2/(Km^2)$ by Lemma 3. □

Lemma 5 (Pilot versus centralized MLE)**.** *Under* **?? 1?? 2,** $\theta_r^{(0)} - \widehat{\theta}_N = \mathbb{E}R_1^J + K^{-1} \sum_k (R_k^J - \mathbb{E}R_k^J) - R_N = O_p(m^{-r} + (m\sqrt{K})^{-1} + N^{-1})$.

Proof. Average Lemma 2 over k: $\theta_r^{(0)} - \theta_0 = K^{-1} \sum_k U_k + K^{-1} \sum_k R_k^J = U_N + \mathbb{E}R_1^J + K^{-1} \sum_k (R_k^J - \mathbb{E}R_k^J)$, since $K^{-1} \sum_k U_k = N^{-1} \sum_{i=1}^N \psi(X_i) = U_N$ (equal block sizes). Subtract $\widehat{\theta}_N - \theta_0 = U_N + R_N$ from (3) applied to the full sample; the influence terms cancel exactly. The orders follow from Lemmas 2 and 4 and $\mathbb{E}\|R_N\|^2 \le CN^{-2}$ via Markov's inequality. □

Lemma 6 (Quadratic contraction of one Newton step)**.** *Under* **??** *3, let* E_N *be the event (of probability tending to one) on which* $\widehat{\theta}_N \in U$ *is an interior zero of* g_N *and the bounds* M_N, Λ_N *hold. On* $E_N \cap \{\vartheta \in U\}$, $\|\mathcal{N}_N(\vartheta) - \widehat{\theta}_N\| \le \frac{1}{2} M_N \Lambda_N \|\vartheta - \widehat{\theta}_N\|^2$.

Proof. Let $e = \vartheta - \widehat{\theta}_N$. Since U is convex, $\widehat{\theta}_N + te \in U$ for $t \in [0,1]$ and $g_N(\vartheta) = g_N(\widehat{\theta}_N) + \int_0^1 H_N(\widehat{\theta}_N + te)\, e\, \mathrm{d}t = \int_0^1 H_N(\widehat{\theta}_N + te)\, e\, \mathrm{d}t$. Hence

$$\mathcal{N}_N(\vartheta) - \widehat{\theta}_N = e - H_N(\vartheta)^{-1} g_N(\vartheta) = H_N(\vartheta)^{-1} \int_0^1 \big[H_N(\vartheta) - H_N(\widehat{\theta}_N + te)\big] e\, \mathrm{d}t.$$

The Lipschitz bound gives $\|H_N(\vartheta) - H_N(\widehat{\theta}_N + te)\| \le \Lambda_N (1-t)\|e\|$, and $\int_0^1 (1-t)\, \mathrm{d}t = 1/2$. □

Lemma 7 (q steps)**.** *Let* $\theta^{(0)} \xrightarrow{p} \theta_0$ *and define* $\theta^{(j+1)} = \mathcal{N}_N(\theta^{(j)})$. *Under* **??** *3, with probability tending to one all iterates* $\theta^{(0)}, \ldots, \theta^{(q)}$ *lie in* U *and* $\|\theta^{(q)} - \widehat{\theta}_N\| \le (\frac{1}{2} M_N \Lambda_N)^{2^q - 1} \|\theta^{(0)} - \widehat{\theta}_N\|^{2^q}$.

Proof. Let $\kappa_N = \frac{1}{2}M_N\Lambda_N = O_p(1)$ and $e_j = \|\theta^{(j)} - \widehat{\theta}_N\|$. Since $e_0 \xrightarrow{p} 0$ and $\widehat{\theta}_N \xrightarrow{p} \theta_0$, there is a ball $B \subset U$ around θ_0 such that $P(\theta^{(0)} \in B, \widehat{\theta}_N \in B) \to 1$. On E_N, induction with Lemma 6 gives $e_{j+1} \le \kappa_N e_j^2$ and, because $\kappa_N e_0 \xrightarrow{p} 0$, $e_{j+1} \le e_j$ eventually, so that each iterate remains in the ball and the recursion applies q times: $e_q \le \kappa_N^{1+2+\cdots+2^{q-1}} e_0^{2^q} = \kappa_N^{2^q-1} e_0^{2^q}$. □

Proof of Theorem 1. By Lemma 5 and $m \to \infty$ (which forces $K = o(N)$), $\theta_r^{(0)} - \theta_0 = U_N + O_p(1) \xrightarrow{p} 0$. Lemma 7 applies and, with $\kappa_N = O_p(1)$,

$$\sqrt{N}\|\theta_r^{(q)} - \widehat{\theta}_N\| = O_p\Big(\sqrt{N}\big[m^{-r} + (m\sqrt{K})^{-1} + N^{-1}\big]^{2^q}\Big) \le O_p\Big(3^{2^q}\sqrt{N}\big[m^{-r2^q} + (m\sqrt{K})^{-2^q} + N^{-2^q}\big]\Big),$$

using $(a+b+c)^{2^q} \le 3^{2^q}\max(a,b,c)^{2^q}$. With $m = N/K$: $\sqrt{N}m^{-r2^q} = N^{1/2}(K/N)^{r2^q} \to 0$ iff $K^{r2^q} = o(N^{r2^q-1/2})$ iff $K = o(N^{1-1/(2^{q+1}r)})$. Next, $\sqrt{N}(m\sqrt{K})^{-2^q} = N^{1/2}(\sqrt{K}/N)^{2^q} = N^{1/2}K^{2^{q-1}}N^{-2^q}$; for $q = 0$ this is $(K/N)^{1/2} \to 0$; for $q \ge 1$ and $K \le N$ it is at most $N^{1/2+2^{q-1}-2^q} = N^{1/2-2^{q-1}} \le N^{-1/2} \to 0$. Finally $\sqrt{N}N^{-2^q} \to 0$. The limit law follows from $\sqrt{N}(\widehat{\theta}_N - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_\psi)$ and Slutsky's lemma. □

Proof of Corollary 1. $\sqrt{N}(\theta_r^{(q)} - \theta_0) \Rightarrow \mathcal{N}(0, \Sigma_\psi)$ and $\widehat{\Sigma}_N \xrightarrow{p} \Sigma_\psi$ give $\sqrt{N}\widehat{\Sigma}_{N,jj}^{-1/2}(\theta_r^{(q,j)} - \theta_{0,j}) \Rightarrow \mathcal{N}(0,1)$ by Slutsky's lemma. Consistency of $-H_N(\theta_r^{(q)})^{-1}$ under correct specification follows from **??** 3 (Lipschitz continuity of H_N on U, $\theta_r^{(q)} \xrightarrow{p} \theta_0$) and $-H_N(\theta_0) \xrightarrow{p} I(\theta_0)$; the sandwich case is analogous with a law of large numbers for the aggregated score outer products. □

Proof of Theorem 3. Let $e = \vartheta - \widehat{\theta}_N$ and $\bar{H} := \int_0^1 H_N(\widehat{\theta}_N + te)\,\mathrm{d}t$, so that $g_N(\vartheta) = \bar{H}e$ on $E_N \cap \{\vartheta \in U\}$. Then

$$\tilde{\vartheta} - \widehat{\theta}_N = e - \widehat{H}^{-1}(\bar{H}e + \delta_g) = \widehat{H}^{-1}\big(\widehat{H} - \bar{H}\big)e - \widehat{H}^{-1}\delta_g,$$

and $\|\widehat{H} - \bar{H}\| \le \|\widehat{H} - H_N(\widehat{\theta}_N)\| + \int_0^1 \|H_N(\widehat{\theta}_N) - H_N(\widehat{\theta}_N + te)\|\mathrm{d}t \le \|\widehat{H} - H_N(\widehat{\theta}_N)\| + \frac{1}{2}\Lambda_N\|e\|$, which gives (10). Iterating with a fixed $\widehat{H}$ (or with a fresh $\widehat{H}$ of the same accuracy at each step) yields $e_{j+1} = O_p(\delta_N e_j + e_j^2)$ and, by induction, $e_q = O_p(e_0\delta_N^q + \sum_{\text{mixed}} \cdots + e_0^{2^q})$ in which every mixed term is dominated by $e_0\delta_N^q$ or $e_0^{2^q}$ once $e_0 \to 0$ and $\delta_N \to 0$; the iterates remain in U as in Lemma 7 because $\delta_N + \kappa_N e_0 \xrightarrow{p} 0$. With $e_0 = O_p(m^{-r} + \sqrt{K}/N + N^{-1})$ the binding condition is $\sqrt{N}m^{-r}\delta_N^q \to 0$ (the terms $\sqrt{N}(\sqrt{K}/N)\delta_N^q$ and $\sqrt{N}N^{-1}\delta_N^q$ vanish for any $\delta_N \le 1$ and $K \le N$), together with the exact-curvature condition $\sqrt{N}e_0^{2^q} \to 0$ of Theorem 1.

(a) Write $\widehat{H} - H_N(\widehat{\theta}_N) = K^{-1}\sum_k [H_{D_k}(\widehat{\theta}_{D_k}) - H_{D_k}(\theta_0)] + [H_N(\theta_0) - H_N(\widehat{\theta}_N)]$. The second bracket is $O_p(\|\widehat{\theta}_N - \theta_0\|) = O_p(N^{-1/2})$ by **??** 3. For the first, **??** 4 gives, for each k and $\Delta_k := \widehat{\theta}_{D_k} - \theta_0$,

$$H_{D_k}(\widehat{\theta}_{D_k}) - H_{D_k}(\theta_0) = \nabla H_{D_k}(\theta_0)[\Delta_k] + r_k, \qquad \|r_k\| \le \tfrac{1}{2}\bar{L}_k\|\Delta_k\|^2,$$

with $\bar{L}_k$ the block average of $L(X_i)$. By Cauchy–Schwarz and **??** 4, $\mathbb{E}\|r_k\| \le \frac{1}{2}(\mathbb{E}\bar{L}_k^2)^{1/2}(\mathbb{E}\|\Delta_k\|^4)^{1/2} = O(m^{-1})$. Write $\nabla H_{D_k}(\theta_0) = T + V_k$ with $T = \mathbb{E}\nabla h(X;\theta_0)$ and V_k a centered block average with

$\mathbb{E}\|V_k\|^2 = O(m^{-1})$, and $\Delta_k = U_k + R_{D_k}$. Then $K^{-1}\sum_k T[U_k] = T[U_N] = O_p(N^{-1/2})$; $K^{-1}\sum_k T[R_{D_k}] = O_p(m^{-1})$ by **??** 1; and $\mathbb{E}\|V_k[U_k]\| \le (\mathbb{E}\|V_k\|^2)^{1/2}(\mathbb{E}\|U_k\|^2)^{1/2} = O(m^{-1})$. Collecting terms, $\delta_N = m^{-1} + N^{-1/2}$, and the condition $\sqrt{N}m^{-r}m^{-q} \to 0$ reads $K = o(N^{1-1/(2(r+q))})$; the $N^{-1/2}$ part of δ_N never binds. Since $r + q \le 2^q r$ for $q \ge 1$, this condition implies that of Theorem 1.

(b) $H_{D_1}(\vartheta) - H_N(\widehat{\theta}_N) = [H_{D_1}(\vartheta) - H_{D_1}(\theta_0)] + [H_{D_1}(\theta_0) - \bar{H}(\theta_0)] + [\bar{H}(\theta_0) - H_N(\theta_0)] + [H_N(\theta_0) - H_N(\widehat{\theta}_N)]$, where $\bar{H}(\theta_0) = \mathbb{E}h(X;\theta_0)$. The first and last brackets are $O_p(\|\vartheta - \theta_0\| + N^{-1/2})$ by the Lipschitz property (applied to H_{D_1}, for which the same bound is assumed as for H_N on U, with probability tending to one); the second is $O_p(m^{-1/2})$ and the third $O_p(N^{-1/2})$ by the central limit theorem for block averages of $h(X_i;\theta_0)$. Hence $\delta_N = m^{-1/2} + \|\vartheta - \theta_0\|$, and along the iteration $\|\theta^{(j)} - \theta_0\| = O_p(N^{-1/2} + e_j)$, so the binding condition is $\sqrt{N}m^{-r}m^{-q/2} \to 0$, i.e. $K = o(N^{1-1/(2r+q)})$.

(c) Rounding each entry of $H_N(\vartheta)$ to b significant bits yields $\widehat{H}$ with $|\widehat{H}_{ab} - H_{N,ab}| \le 2^{-b}|H_{N,ab}|$, hence $\|\widehat{H} - H_N(\vartheta)\| \le p\,2^{-b}\|H_N(\vartheta)\|_{\max}$, and $\|H_N(\vartheta) - H_N(\widehat{\theta}_N)\| \le \Lambda_N\|e\|$; so $\delta_N = O_p(\varepsilon) + O_p(\|e\|)$, the second part being absorbed in the $\|e\|^2$ term. Likewise $\|\delta_g\| \le p\,\varepsilon\,\|g_N(\vartheta)\|_{\max} \le p\,\varepsilon\,\|\bar{H}\|\|e\| = O_p(\varepsilon\|e\|)$. Inserting into (10) with $\|e_0\| = O_p(m^{-r} + \sqrt{K}/N + N^{-1})$, first-order equivalence requires $\sqrt{N}\varepsilon m^{-r} \to 0$ (the other two terms vanish for $\varepsilon \le 1$). With $K \sim N^\alpha$, $m^{-r} = N^{-r(1-\alpha)}$ up to constants, so $\sqrt{N}\varepsilon m^{-r} \to 0$ iff $\log_2(1/\varepsilon) = b \ge (\frac{1}{2} - r(1-\alpha))\log_2 N + \omega(1)$ when $\frac{1}{2} - r(1-\alpha) \ge 0$, i.e. $\alpha \ge 1 - 1/(2r)$; otherwise $\sqrt{N}m^{-r} \to 0$ already and any $b \ge 0$ works. □

Proof of Proposition 2. (i) With weights m_k/N, $\sum_k (m_k/N)U_k = N^{-1}\sum_i \psi(X_i) = U_N$, so Lemma 5 holds with $\mathbb{E}R^J$ replaced by $\sum_k (m_k/N)\mathbb{E}R_k^J$, of norm at most $C_b(\sum_j |w_j|s_j^r)N^{-1}\sum_k m_k^{1-r} = C_b(\cdots)m_{\text{eff}}^{-r}$, and with centered second moment $\sum_k (m_k/N)^2 CA_r^2/m_k^2 = CA_r^2 K/N^2$. The Newton steps use $g_N = \sum_k (m_k/N)g_{D_k}$ and $H_N = \sum_k (m_k/N)H_{D_k}$, which are the full-sample derivatives, so Lemmas 6 and 7 are unchanged. With equal weights, $K^{-1}\sum_k U_k - U_N = \sum_i c_i\psi(X_i)$ with $c_i = (Km_{k(i)})^{-1} - N^{-1}$; its covariance is $(\sum_i c_i^2)\Sigma_\psi$ and $\sum_i c_i^2 = \sum_k m_k[(Km_k)^{-1} - N^{-1}]^2 = K^{-2}\sum_k m_k^{-1} - N^{-1}$, which is ≥ 0 with equality iff all m_k are equal (Jensen); when the sizes are bounded away from equality in the sense $K^{-1}\sum_k (N/(Km_k)) \ge 1 + \eta$, this is $\ge \eta/N$, of exact order N^{-1}, so the discrepancy is of exact order $N^{-1/2}$ and $\sqrt{N}(\theta^{(0)} - \widehat{\theta}_N)$ does not vanish. One Newton step squares it to $O_p(N^{-1})$ by Lemma 6. (ii) The only properties of ψ used in Lemmas 1 to 5 are $\mathbb{E}\psi = 0$ and that the same ψ appears in (3) for every sub-sample; **??** 3 is stated for the pseudo-MLE with the sandwich limit. The proofs go through verbatim. □

B. Proofs for the exponential model

Let $Y = \theta_0\bar{X}_n$, so that $nY \sim \text{Gamma}(n,1)$, $\widehat{\theta}_n = \theta_0/Y$ and $R_n = \theta_0(1-Y)^2/Y = \theta_0(Y^{-1} - 2 + Y)$. The Gamma moments $\mathbb{E}(nY)^{-k} = \Gamma(n-k)/\Gamma(n)$ $(k < n)$ give $\mathbb{E}Y = 1$, $\mathbb{E}Y^2 = 1 + 1/n$, $\mathbb{E}Y^{-1} = n/(n-1)$ and $\mathbb{E}Y^{-2} = n^2/((n-1)(n-2))$.

Proof of Theorem 2(i). $\mathbb{E}\widehat{\theta}_n = \theta_0 \mathbb{E}Y^{-1} = \theta_0 n/(n-1)$ and $\operatorname{Var}\widehat{\theta}_n = \theta_0^2(\mathbb{E}Y^{-2} - (\mathbb{E}Y^{-1})^2) = \theta_0^2 n^2/((n-1)^2(n-2))$. Next $\mathbb{E}R_n = \theta_0(n/(n-1) - 2 + 1) = \theta_0/(n-1) = \theta_0 \sum_{a\geq 1} n^{-a}$, so $b_a = \theta_0$ for all a and the remainder $\rho_n = \theta_0 \sum_{a\geq r} n^{-a}$ satisfies $\|\rho_n\| \leq 2\theta_0 n^{-r}$ for $n \geq 2$. Expanding $(Y^{-1} - 2 + Y)^2 = Y^{-2} - 4Y^{-1} + 6 - 4Y + Y^2$,

$$\mathbb{E}R_n^2 = \theta_0^2 \Big[\frac{n^2}{(n-1)(n-2)} - \frac{4n}{n-1} + 3 + \frac{1}{n} \Big] = \theta_0^2 \frac{3n+2}{n(n-1)(n-2)},$$

and $n^2 \mathbb{E}R_n^2/\theta_0^2 = n(3n+2)/((n-1)(n-2))$ decreases from $33/2$ at $n=3$ to 3, which gives $C = 33\theta_0^2/2$ with $n_0 = 3$. The influence function $\psi(x) = \theta_0 - \theta_0^2 x$ is the same for all sub-samples and $\Sigma_\psi = \theta_0^2 = I(\theta_0)^{-1}$. □

Proof of (ii). A sub-block of level j has $n = m/s_j$ observations, so its MLE has mean $\theta_0 n/(n-1) = \theta_0 m/(m-s_j)$. Hence $\mathbb{E}J_{r,k} = \theta_0 \sum_j w_j m/(m-s_j)$ and $\mathbb{E}J_{r,k} - \theta_0 = \theta_0 \sum_j w_j [m/(m-s_j) - 1] = \theta_0 \sum_j w_j s_j/(m-s_j)$, using $\sum_j w_j = 1$. Expanding $s_j/(m-s_j) = \sum_{a\geq 1} s_j^a m^{-a}$ and using (5), the terms $a < r$ vanish and the leading term is $\theta_0 m^{-r} \sum_j w_j s_j^r = \beta_r \theta_0 m^{-r}$. For $s = (1,2)$: $\theta_0[2/(m-1) - 2/(m-2)] = -2\theta_0/((m-1)(m-2))$. For $(1,2,4)$: $\theta_0[(8/3)/(m-1) - 4/(m-2) + (4/3)/(m-4)] = 8\theta_0/((m-1)(m-2)(m-4))$. Both identities, and $\beta_2 = -2$, $\beta_3 = 8$, $\beta_4 = -64$, are verified in rational arithmetic in `theory/symbolic_checks.py`. □

Proof of (iii). Let $h = m/2$ and S_A, S_B be the sums over the two halves, $S = S_A + S_B \sim \text{Gamma}(m, \theta_0)$ and $W = S_A/S \sim \text{Beta}(h,h)$, independent of S. Then $\widehat{\theta}_{D_k} = m/S$, $\widehat{\theta}_A = h/(SW)$, $\widehat{\theta}_B = h/(S(1-W))$ and $J_{2,k} = S^{-1}G(W)$ with $G(W) = 2m - \frac{h}{2}[W^{-1} + (1-W)^{-1}]$. Independence gives $\mathbb{E}J_{2,k} = \mathbb{E}S^{-1}\,\mathbb{E}G(W)$ and $\mathbb{E}J_{2,k}^2 = \mathbb{E}S^{-2}\,\mathbb{E}G(W)^2$, with $\mathbb{E}S^{-k} = \theta_0^k \Gamma(m-k)/\Gamma(m)$ and the Beta inverse moments $\mathbb{E}W^{-a}(1-W)^{-c} = \Gamma(h-a)\Gamma(h-c)\Gamma(2h)/(\Gamma(h)^2\Gamma(2h-a-c))$ for $a, c < h$. Simplifying the resulting rational functions (SymPy, check S6) gives the stated $\operatorname{Var} J_{2,k}$ and the expansion $\theta_0^2(m^{-1} + 4m^{-2} + O(m^{-3}))$; the same expansion holds for $\operatorname{Var}\widehat{\theta}_m = \theta_0^2 m^2/((m-1)^2(m-2))$. The condition $m > 4$ ensures $h > 2$ so that all inverse moments exist. □

Proof of (iv). With $g_N(x) = 1/x - \bar{X}_N$ and $H_N(x) = -1/x^2$, $\mathcal{N}_N(x) = x + x^2(1/x - \bar{X}_N) = 2x - x^2\bar{X}_N$. Since $\widehat{\theta}_N = 1/\bar{X}_N$, $\mathcal{N}_N(x) - \widehat{\theta}_N = -\bar{X}_N x^2 + 2x - 1/\bar{X}_N = -\bar{X}_N(x - 1/\bar{X}_N)^2 = -(x - \widehat{\theta}_N)^2/\widehat{\theta}_N$. Writing $e_j = \theta^{(j)} - \widehat{\theta}_N$ gives $e_{j+1} = -e_j^2/\widehat{\theta}_N$ and, by induction, $e_q = -e_0^{2^q}/\widehat{\theta}_N^{2^q - 1}$. In units of $\widehat{\theta}_N$, $u_{j+1} = -u_j^2$ with $u_j = e_j/\widehat{\theta}_N$, so $|u_j| \to 0$ iff $|u_0| < 1$, i.e. $0 < x < 2\widehat{\theta}_N$; $|u_0| > 1$ gives $|u_j| \to \infty$ with $u_j < 0$ for $j \geq 1$; $u_0 = \pm 1$ gives $u_1 = -1$ and then $u_j = -1$, i.e. the iterate 0, where the log-likelihood is undefined. □

Proof of (v)–(vi). Let $\alpha = \alpha(r,q)$ and $K/N^\alpha \to c \in (0, \infty]$, with $m \to \infty$. By Lemma 5 and (ii), $\theta_r^{(0)} - \widehat{\theta}_N = \mathbb{E}R_1^J + \bar{\xi}_N - R_N$ with $\mathbb{E}R_1^J = \beta_r \theta_0 m^{-r}(1 + O(m^{-1}))$, $\bar{\xi}_N = O_p(\sqrt{K}/N)$ and $R_N =$

$O_p(N^{-1})$; part (i) supplies the constants. Since $m^{-r} = (K/N)^r$ and $r(1-\alpha) = 2^{-(q+1)}$,

$$N^{1/2^{q+1}} m^{-r} = N^{1/2^{q+1}} (K/N^{\alpha})^r N^{-r(1-\alpha)} = (K/N^{\alpha})^r \to c^r,$$

while $N^{1/2^{q+1}} \sqrt{K}/N \le N^{1/2^{q+1}+\alpha/2-1} \to 0$ because $\alpha < 1$ and $1/2^{q+1} \le 1/2$ give $1/2^{q+1} + \alpha/2 - 1 < 0$; likewise $N^{1/2^{q+1}} N^{-1} \to 0$. Hence $N^{1/2^{q+1}} (\theta_r^{(0)} - \widehat{\theta}_N) \xrightarrow{p} \beta_r \theta_0 c^r$ when $c < \infty$ and $\xrightarrow{p} \pm\infty$ when $c = \infty$. For $q = 0$, $N^{1/2^{q+1}} = \sqrt{N}$ and the claim follows. For $q \ge 1$, (iv) gives $\sqrt{N}(\theta_r^{(q)} - \widehat{\theta}_N) = -\sqrt{N}(\theta_r^{(0)} - \widehat{\theta}_N)^{2^q} / \widehat{\theta}_N^{2^q-1} = -[N^{1/2^{q+1}} (\theta_r^{(0)} - \widehat{\theta}_N)]^{2^q} / \widehat{\theta}_N^{2^q-1}$, and $\widehat{\theta}_N \xrightarrow{p} \theta_0$ with the continuous mapping theorem yields $-\theta_0 (\beta_r c^r)^{2^q}$ when $c < \infty$ and divergence to $-\infty$ in probability when $c = \infty$. Combining with $\sqrt{N}(\widehat{\theta}_N - \theta_0) \Rightarrow \mathcal{N}(0, \theta_0^2)$ proves the displayed limit law for $(2,1)$. Strictly, (iv) requires the pilot to be positive for the iteration to be meaningful; $P(\theta_r^{(0)} \le 0) \to 0$ since $\theta_r^{(0)} \xrightarrow{p} \theta_0 > 0$. □

C. Reference implementations

The replication package (Section D) contains the package `pamle` with the following components, all exercised by the test suite.

- `jackknife.py`: exact rational Vandermonde weights for any pattern, the constants β_r and A_r, condition numbers, and the combination of level averages.
- `likelihoods.py`: exponential, logistic and Poisson models with batched Newton fitting over blocks and sub-blocks (step halving, convergence on the score, status codes for non-convergence, singular curvature and divergence, and a separation safeguard for logistic regression that refuses a stationary point with fitted $|\eta| > 30$ or a numerically singular information matrix). Derivatives are checked against finite differences.
- `algorithms.py`: local split fits, corrected worker estimates, size-weighted pilots, global Newton corrections with exact, stale, single-machine and quantized curvature (linear solves, condition numbers and step norms as diagnostics; no explicit inverse in the Newton step), the aggregated estimating-equation estimator of Lin and Xi (2011), the surrogate-likelihood estimator of Jordan et al. (2019) with its full local maximization, and the analytic first-order bias correction of canonical GLMs (Cox and Snell, 1968; Cordeiro and McCullagh, 1991).
- `communication.py`: rounds, messages, numbers and bytes for every variant of Table 4.
- `protocol.py`, `worker.py`: the coordinator–worker protocol over TCP with length-prefixed messages, per-stage wall-clock, per-worker compute times, latency percentiles and byte counts (Algorithm 1).

Analytic bias correction used as a baseline. For a canonical-link GLM with per-observation score $\psi_i(\beta) = x_i(y_i - \mu_i)$, the $O(1/n)$ bias of the MLE conditional on the design is b_1/n with

$$b_1 = \tfrac{1}{2} I^{-1} \left[n^{-1} \textstyle\sum_i \nabla^2 \psi_i [I^{-1}] \right] = -\tfrac{1}{2} I^{-1} n^{-1} \textstyle\sum_i x_i \mu''(\eta_i) x_i^\top I^{-1} x_i,$$

where $I = n^{-1} \sum_i x_i x_i^\top \mu'(\eta_i)$; this is the form of Cordeiro and McCullagh (1991) and reduces to $b_1 = \theta_0$ in the exponential model. The corrected local estimate $\widehat{\theta}_{D_k} - \hat{b}_1/m$ needs one local fit and no sub-blocks; it is the local analogue of the debiased weighted estimator of Gu and Chen (2023) and provides, in Section 9, a baseline that attains the order-2 exponent at one third of the local fits of the half-block jackknife whenever the model's bias formula is available and correct.

D. Replication

Everything reported is regenerated from committed raw outputs by the scripts in `experiments/`; no number in the tables or figures is typed by hand. The directory `references/initial_pamle/` contains the prior draft, its archive and SHA-256 checksums; its Monte Carlo (seed 20261028, $N = 2^{20}$, $m \in \{256, 64, 16\}$, 500 replications) was re-run and matched to 5×10^{-13} before any change was made, and the test `tests/test_reproducibility.py` reproduces its summary table from the new code.

Commands.

```
python3 -m pip install -r requirements.txt
pytest -q                                    # unit tests, exact identities, protocol
python3 theory/symbolic_checks.py            # SymPy verification of all identities
python3 experiments/run_monte_carlo.py --config experiments/configs/main_exponential.yam
python3 experiments/scatter_cluster.py --configs experiments/configs/main_logistic.yaml
        experiments/configs/main_poisson.yaml experiments/configs/robustness.yaml --laun
python3 experiments/run_boundary.py; python3 experiments/run_ablations.py --model logist
python3 experiments/run_skewed.py
python3 experiments/run_benchmarks.py --config experiments/configs/benchmarks.yaml \
        --out results/raw/benchmarks.jsonl   # requires the cluster
python3 experiments/process_results.py --figures; python3 experiments/make_tables.py
bash scripts/compile_paper.sh; bash scripts/validate_pdf.sh
```

Seeds are derived from `seed_base = 20261028` and a hash of the cell description; every raw file has a JSON manifest with git commit, Python and NumPy versions, host, wall time and the SHA-256 of the output (`experiments/EXPERIMENT_MANIFEST.csv`). Monte Carlo cells were

run on the laptop (Apple M5 Max, Python 3.14.4, NumPy 2.4.4) and on 18 cluster nodes (Python 3.9.6/3.14, NumPy 2.0.2/2.5.4); floating-point results differ across these environments at the level of 10^{-12} relative, far below the Monte Carlo standard errors reported.