Local Bias Correction versus Global Synchronization in Distributed Maximum Likelihood

Rate Frontiers, Exact Exponential Analysis, and Reproducible Monte Carlo Evidence

Simon-Pierre Boucher[1] **and** **GPT-6.1**[2]

[1] Université du Québec en Outaouais [2] AI co-author designation

Working paper • October 28, 2026 • Theoretical and numerical research note

Abstract

Distributed maximum likelihood estimation (MLE) trades statistical efficiency against network communication and local computation. We analyze a two-phase construction that combines a within-worker split-sample jackknife with a single globally aggregated Newton step. Under fixed-dimensional asymptotic linearity, a second-order bias expansion and local Hessian regularity, we prove first-order equivalence to centralized MLE for $K = o(N^{7/8})$ partitions; the proof explicitly retains the jackknife's centered random remainder. A higher-order local extrapolation yields $K = o(N^{11/12})$ under stronger assumptions. For exponential observations, exact bias formulas and a Newton identity establish a nonzero limiting shift at the $N^{7/8}$ boundary of the specified algorithm. Reproducible exact-Gamma Monte Carlo experiments confirm the qualitative pattern. We account for derivative-communication costs and compare extra local estimation with an additional Newton synchronization. These results provide a transparent theoretical and numerical working note, not an established claim of methodological priority or improved wall-clock performance.

Keywords: maximum likelihood; distributed econometrics; split-sample jackknife; one-step Newton; asymptotic equivalence; communication complexity; Monte Carlo.

JEL codes: C13, C15, C55, C63.

Disclosure. The second author label is an AI designation supplied for this draft. The mathematics is conditional on explicit assumptions, and the numerical experiment is executable and independently reproducible. Researcher review, literature-priority checks, and empirical performance benchmarking remain necessary prior to submission.

Contents

1 Introduction **3**

2 Model, estimators, and information exchange **3**

3 An exact decomposition of local bias correction **5**

4 One Newton step and the $7/8$ frontier **6**

5 Exact exponential-model analysis **8**

6 Higher-order local extrapolation **9**

7 Numerical experiment: exact Gamma simulation **11**

8 Communication and computation accounting **13**

9 Limitations and research agenda **14**

10 Conclusion **14**

Statements and reproducibility **15**

A Proof details for the exponential remainder **15**

B Exact bias of the quarter-block extrapolation **16**

1. Introduction

Massive datasets distributed across geographically or operationally separate devices create a tension between three objectives: statistically efficient estimation, low network communication, and computational scalability. For a regular parametric model the centralized maximum likelihood estimator can attain the inverse Fisher-information asymptotic variance, but collecting all observations on one machine may be expensive or prohibited. An alternative is to fit the model independently on K data partitions and average the local estimators. This does not automatically preserve centralized statistical efficiency: nonlinear estimation induces a local bias that does not disappear merely by averaging independent estimates.

Early systematic analyses of divide-and-conquer statistical optimization include Zhang et al. (2013), who study averaging and communication-efficient alternatives, while Huang and Huo (2019) develop a distributed one-step estimator. Multi-round and adaptive distributed procedures are represented by Fan et al. (2023). Separately, jackknife extrapolation has a long history in bias correction; Dhaene and Jochmans (2015) treat split-panel corrections, including higher-order variants, in nonlinear fixed-effects models. Our starting point is to combine these familiar elements and quantify the *exchange rate* between extra local statistical work and an additional global synchronization.

We focus on a sharply defined question: for fixed parameter dimension p, how fast may the partition count K grow with the global sample size N while a locally bias-corrected, globally one-step estimator remains first-order equivalent to the centralized MLE? For the half-block jackknife, we establish $K = o(N^{7/8})$ under explicit regularity conditions. Our proof takes special care to retain a random term of order $1/(m\sqrt{K})$, where $m = N/K$ is the local sample size. Omitting that term would produce an incomplete argument, even if the final exponent happened to be correct.

Our contributions are deliberately bounded. We give a transparent sufficient-rate theorem, an exact verification and boundary distribution for an exponential model, a higher-order generalization, and a reproducible computational illustration. We do *not* establish that this procedure dominates the best available distributed estimators, nor do we prove a universal communication lower bound. The asymptotic regime is fixed-p, i.i.d., correctly specified, and locally regular. The exact exponential model actually admits trivial sufficient-statistic aggregation; it is therefore a proof-of-concept and counterexample laboratory, not a favorable application of distributed approximate MLE.

2. Model, estimators, and information exchange

2.1 Likelihood and partitioning

Let $X_1, \dots, X_N$ be i.i.d. observations with density $f(x;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$, and true value θ_0 in the interior of Θ. Let K divide N, set $m = N/K$, and partition the index set into disjoint blocks $\mathcal{D}_1, \dots, \mathcal{D}_K$, each of size m. The average log-likelihood and centralized MLE are

$$L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log f(X_i;\theta), \qquad \widehat{\theta}_N \in \arg\max_{\theta\in\Theta} L_N(\theta). \tag{1}$$

Worker k computes a local MLE $\widehat{\theta}_{k,m}$ based on $\mathcal{D}_k$. In the baseline construction m is even and each block has two disjoint halves A_k, B_k, with half-sample MLEs $\widehat{\theta}_{k,A}$ and $\widehat{\theta}_{k,B}$. Write

$$\bar{\theta}_{\text{raw}} = K^{-1} \sum_{k=1}^{K} \widehat{\theta}_{k,m}, \tag{2}$$

$$J_k = 2\widehat{\theta}_{k,m} - \tfrac{1}{2}(\widehat{\theta}_{k,A} + \widehat{\theta}_{k,B}), \tag{3}$$

$$\theta^{(0)} = K^{-1} \sum_{k=1}^{K} J_k. \tag{4}$$

This local operation needs three likelihood fits per block but, because all three fits access locally stored observations, introduces no within-cluster data transmission beyond the final parameter vector.

Define the full score and Hessian by $g_N(\theta) = \nabla_\theta L_N(\theta)$ and $H_N(\theta) = \nabla^2_\theta L_N(\theta)$. After the coordinator broadcasts $\theta^{(0)}$, each worker sends its contributions to $g_N(\theta^{(0)})$ and $H_N(\theta^{(0)})$. The coordinator computes

$$\boxed{\theta^{(1)} = \theta^{(0)} - H_N(\theta^{(0)})^{-1} g_N(\theta^{(0)}) \quad .} \tag{5}$$

An invertible Hessian and well-defined, consistent local solutions are assumed throughout; a production implementation may require regularization and step damping, neither of which is covered by the present theory.

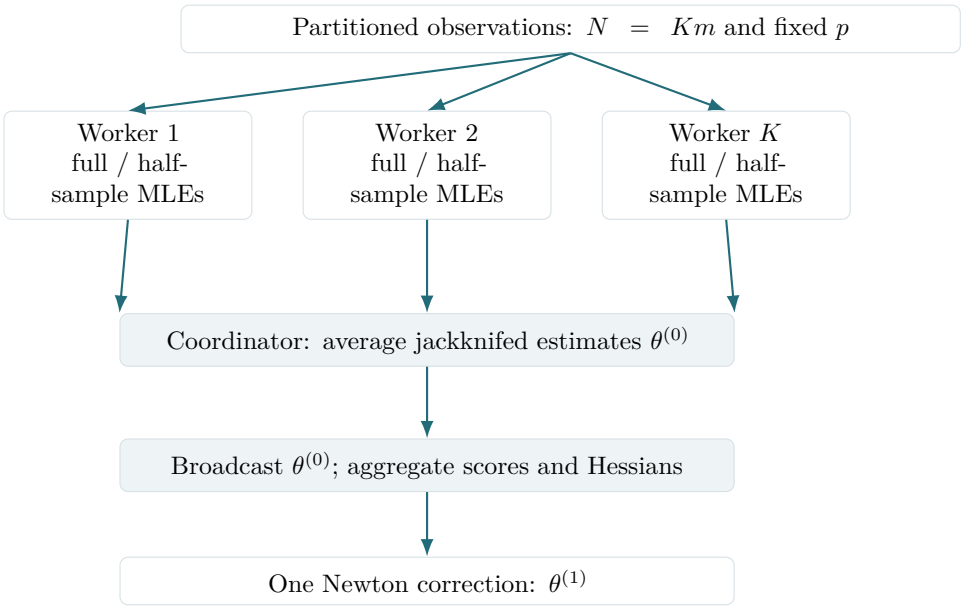


Figure 1: Local bias correction followed by one globally aggregated Newton step. Arrows indicate logical phases, not measured network latency.

2.2 Regularity and the benchmark MLE

Our central statistical result is conditional upon the following strong but explicit assumptions, stated for a *sequence* with $N \to \infty$ and $K = K_N \to \infty$ permitted.

Assumption 2.1 (Uniform asymptotic linear expansion)**.** For every relevant sub-sample size $n \in \{m, m/2, N\}$, its MLE has the representation

$$\widehat{\theta}_n - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i) + R_n, \tag{6}$$

where $\mathbb{E}\psi(X_i) = 0$, $\mathbb{E}\|\psi(X_i)\|^2 < \infty$, $\mathrm{Var}(\psi(X_i)) = \mathcal{I}(\theta_0)^{-1}$ for a positive-definite Fisher information matrix, and a common constant $C < \infty$ satisfies

$$\mathbb{E}\|R_n\|^2 \le Cn^{-2}. \tag{7}$$

For correctly specified, regular parametric MLE, $\psi(x) = \mathcal{I}(\theta_0)^{-1}\nabla_\theta \log f(x;\theta_0)$. The expansion is required jointly with the same ψ on all nested subsamples.

Assumption 2.2 (Second-order bias expansion)**.** There exist vectors b_1, b_2 and a constant C_b, not depending on n, for which

$$\mathbb{E}R_n = \frac{b_1}{n} + \frac{b_2}{n^2} + \rho_n, \qquad \|\rho_n\| \le C_b n^{-3}, \tag{8}$$

for the sample sizes involved. The existence of this expansion is not implied by mere consistency or asymptotic normality.

Assumption 2.3 (Local Newton regularity)**.** With probability tending to one, $\widehat{\theta}_N$ is a consistent interior stationary maximizer, $g_N(\widehat{\theta}_N) = 0$, and there is a fixed convex neighborhood $\mathcal{U}$ of θ_0 such that

$$\sup_{\theta\in\mathcal{U}} \|H_N(\theta)^{-1}\| = O_p(1), \qquad \sup_{u\neq v\in\mathcal{U}} \frac{\|H_N(u) - H_N(v)\|}{\|u - v\|} = O_p(1). \tag{9}$$

The probability that both $\widehat{\theta}_N$ and the initializer lie in $\mathcal{U}$ tends to one. The centralized MLE obeys $\sqrt{N}(\widehat{\theta}_N - \theta_0) \Rightarrow \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$.

Remark 2.4. The assumptions deliberately fix p and require i.i.d. partitions, finite moments of inverse-information-type quantities, and a well-behaved local optimizer. Weak identification, incidental parameters, dependent samples, and growing p require different analysis. The theorem does not apply merely because code can execute on multiple machines.

3. An exact decomposition of local bias correction

Proposition 3.1 (Half-block jackknife initializer)**.** *Under Theorems 2.1 and 2.2, with independent blocks and $m = N/K$ even,*

$$\mathbb{E}[J_k] - \theta_0 = -\frac{2b_2}{m^2} + O(m^{-3}), \tag{10}$$

$$\theta^{(0)} - \widehat{\theta}_N = O_p\left(m^{-2} + \frac{1}{m\sqrt{K}} + N^{-1}\right) \tag{11}$$

$$= O_p\left(\frac{K^2}{N^2} + \frac{\sqrt{K}}{N} + \frac{1}{N}\right). \tag{12}$$

Proof. Write $U_k = m^{-1}\sum_{i\in\mathcal{D}_k}\psi(X_i)$ and define $U_{k,A}$ and $U_{k,B}$ analogously. Because the halves partition $\mathcal{D}_k$ exactly,

$$\tfrac{1}{2}(U_{k,A}+U_{k,B}) = U_k. \tag{13}$$

Applying Equation (6) to the three local estimators yields

$$J_k - \theta_0 = U_k + R_k^J, \qquad R_k^J = 2R_{k,m} - \tfrac{1}{2}(R_{k,A,m/2} + R_{k,B,m/2}). \tag{14}$$

Using Equation (8) at m and $m/2$ gives

$$\begin{aligned}\mathbb{E}R_k^J &= 2\left(\frac{b_1}{m}+\frac{b_2}{m^2}\right) - \left(\frac{2b_1}{m}+\frac{4b_2}{m^2}\right) + O(m^{-3}) \\ &= -\frac{2b_2}{m^2} + O(m^{-3}).\end{aligned}$$

The L^2 triangle inequality and Equation (7) imply

$$\begin{aligned}(\mathbb{E}\|R_k^J\|^2)^{1/2} &\le 2\sqrt{C}/m + \frac{1}{2}(2\sqrt{C}/m + 2\sqrt{C}/m) \\ &= 4\sqrt{C}/m.\end{aligned}$$

Hence $\mathbb{E}\|R_k^J\|^2 \le 16C/m^2$. Independence of blocks eliminates cross terms for centered remainders:

$$\mathbb{E}\left\|K^{-1}\sum_{k=1}^{K}(R_k^J - \mathbb{E}R_k^J)\right\|^2 \le \frac{16C}{Km^2}. \tag{15}$$

Thus, by Markov's inequality,

$$\theta^{(0)} - \theta_0 = \frac{1}{N}\sum_{i=1}^{N}\psi(X_i) - \frac{2b_2}{m^2} + O(m^{-3}) + O_p\left(\frac{1}{m\sqrt{K}}\right). \tag{16}$$

The full-sample expansion gives $\widehat{\theta}_N - \theta_0 = N^{-1}\sum_i \psi(X_i) + O_p(N^{-1})$. Subtraction cancels the leading influence term *exactly*, proving the second assertion. □

Remark 3.2 (Why the random remainder matters)**.** A small *mean* remainder does not imply a small realized remainder. The term $1/(m\sqrt{K}) = \sqrt{K}/N$ in Equation (12) describes variation of centered local second-order terms. Bias cancellation alone would not justify removing it. Averaging over workers does remove part of its magnitude, but only after an explicit second-moment argument.

4. One Newton step and the 7/8 frontier

Lemma 4.1 (Quadratic local error contraction)**.** *Under Theorem 2.3, for any initializer ϑ_N satisfying $\vartheta_N \to_p \theta_0$ and lying with $\widehat{\theta}_N$ in the regularity neighborhood, its undamped Newton update $\mathcal{N}_N(\vartheta_N) = \vartheta_N - H_N(\vartheta_N)^{-1}g_N(\vartheta_N)$ obeys*

$$\|\mathcal{N}_N(\vartheta_N) - \widehat{\theta}_N\| = O_p(\|\vartheta_N - \widehat{\theta}_N\|^2). \tag{17}$$

Proof. Let $e = \vartheta_N - \widehat{\theta}_N$. By the fundamental theorem of calculus, $g_N(\vartheta_N) = \int_0^1 H_N(\widehat{\theta}_N + te)e\,dt$. Since $g_N(\widehat{\theta}_N) = 0$, one obtains

$$\mathcal{N}_N(\vartheta_N) - \widehat{\theta}_N = H_N(\vartheta_N)^{-1} \int_0^1 [H_N(\vartheta_N) - H_N(\widehat{\theta}_N + te)]e\,dt.$$

The local Lipschitz property bounds the integrand by $\Lambda_N(1-t)\|e\|^2$. Taking norms and integrating yields

$$\|\mathcal{N}_N(\vartheta_N) - \widehat{\theta}_N\| \le \tfrac{1}{2}\|H_N(\vartheta_N)^{-1}\|\Lambda_N\|e\|^2,$$

where the multiplicative factor is $O_p(1)$ by Theorem 2.3. □

Theorem 4.2 (Jackknife–Newton equivalence)**.** *Under Theorems 2.1 to 2.3, with $K = o(N^{7/8})$ and $m = N/K$ even, the estimator in Equations (3) to (5) satisfies*

$$\boxed{\sqrt{N}(\theta^{(1)} - \widehat{\theta}_N) \xrightarrow{p} 0.} \tag{18}$$

Consequently,

$$\sqrt{N}(\theta^{(1)} - \theta_0) \Rightarrow \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right). \tag{19}$$

Proof. By Theorem 3.1, the starting discrepancy is

$$e_0 = \theta^{(0)} - \widehat{\theta}_N = O_p(m^{-2} + m^{-1}K^{-1/2} + N^{-1}).$$

The assumed K regime implies $m \to \infty$ and hence consistency of the initializer, permitting Theorem 4.1. Then

$$\begin{aligned}\|\theta^{(1)} - \widehat{\theta}_N\| &= O_p\left(m^{-4} + \frac{1}{Km^2} + N^{-2}\right) \\ &= O_p\left(\frac{K^4}{N^4} + \frac{K}{N^2} + \frac{1}{N^2}\right).\end{aligned} \tag{20}$$

Multiplication by $\sqrt{N}$ gives

$$\sqrt{N}\|\theta^{(1)} - \widehat{\theta}_N\| = O_p\left(\frac{K^4}{N^{7/2}} + \frac{K}{N^{3/2}} + N^{-3/2}\right). \tag{21}$$

All three deterministic orders vanish whenever $K = o(N^{7/8})$. The first result follows; the limiting distribution then follows from the centralized MLE central limit theorem and Slutsky's lemma. □

Corollary 4.3 (Comparison with an uncorrected start)**.** *If the ordinary local average instead has bias $O_p(m^{-1})$ and a linear expansion of the form used above, then one global Newton step has a sufficient statistical-equivalence regime $K = o(N^{3/4})$. Two global Newton steps, started from the same raw average, have a sufficient regime $K = o(N^{7/8})$ under analogous local regularity at successive steps.*

Proof. The raw starting discrepancy is $O_p(m^{-1} + m^{-1}K^{-1/2} + N^{-1})$ and is dominated by m^{-1} in the relevant regime. One squaring requires $\sqrt{N}m^{-2} \to 0$, or $K = o(N^{3/4})$; two squarings require $\sqrt{N}m^{-4} \to 0$, or $K = o(N^{7/8})$. All stochastic remainders are negligible under these regimes. □

Remark 4.4 (Interpretation rather than priority claim)**.** Theorem 4.2 does not by itself establish a new statistical estimator relative to all prior literature. The well-studied ingredients are local estimator combination, one-step distributed optimization, and split-sample bias correction (Zhang et al., 2013; Huang and Huo, 2019; Dhaene and Jochmans, 2015). Its role here is to establish a self-contained rate calculation and a specific communication–computation comparison suitable for further research.

5. Exact exponential-model analysis

5.1 An analytically tractable likelihood

Suppose $X_i \overset{\text{iid}}{\sim} \text{Exp}(\theta_0)$ using the *rate* parameterization. The average log-likelihood is

$$L_n(\theta) = \log\theta - \theta\bar{X}_n, \qquad \widehat{\theta}_n = \bar{X}_n^{-1}. \tag{22}$$

Let $S_n = \sum_{i=1}^n X_i$, so $S_n \sim \text{Gamma}(n, \text{rate} = \theta_0)$. The Gamma inverse moments give, for $n > 2$,

$$\mathbb{E}[\widehat{\theta}_n] = \theta_0 \frac{n}{n-1}, \tag{23}$$

$$\text{Var}(\widehat{\theta}_n) = \theta_0^2 \frac{n^2}{(n-1)^2(n-2)}. \tag{24}$$

Here $\psi(X) = \theta_0 - \theta_0^2 X$. The remainder has the *exact* form

$$R_n = \widehat{\theta}_n - \theta_0 - \frac{1}{n}\sum_i \psi(X_i) = \frac{(1-\theta_0\bar{X}_n)^2}{\bar{X}_n}, \tag{25}$$

whose moments are

$$\mathbb{E}R_n = \frac{\theta_0}{n-1} = \theta_0(n^{-1} + n^{-2} + O(n^{-3})), \tag{26}$$

$$\mathbb{E}R_n^2 = \theta_0^2 \frac{3n+2}{n(n-1)(n-2)} = O(n^{-2}). \tag{27}$$

These identities verify the key strong mean-square and bias assumptions for this example, with $b_1 = b_2 = \theta_0$. The score and Hessian are $g_N(\theta) = 1/\theta - \bar{X}_N$ and $H_N(\theta) = -1/\theta^2$; local Newton regularity holds around every fixed $\theta_0 > 0$.

5.2 The exact jackknife bias

The local full and half estimators have means $\theta_0 m/(m-1)$ and $\theta_0 m/(m-2)$ respectively. Therefore

$$\mathbb{E}J_k = 2\theta_0 \frac{m}{m-1} - \theta_0 \frac{m}{m-2},$$

$$\boxed{\mathbb{E}J_k - \theta_0 = -\frac{2\theta_0}{(m-1)(m-2)}.} \tag{28}$$

The identity confirms the m^{-2} order without relying on an asymptotic expansion. The split estimates are dependent because they share observations with the full estimate; accounting for this dependence is essential when evaluating variance. Nevertheless, Equation (25) and the decomposition in Equation (14) ensure $\text{Var}(J_k) = \theta_0^2/m + O(m^{-2})$ as $m \to \infty$.

5.3 Exact Newton identity and boundary distribution

A Newton step started from x simplifies to

$$\mathcal{N}_N(x) = 2x - x^2\bar{X}_N. \tag{29}$$

Because $\widehat{\theta}_N = 1/\bar{X}_N$, the following error formula holds *exactly*:

$$\boxed{\mathcal{N}_N(x) - \widehat{\theta}_N = -\frac{(x-\widehat{\theta}_N)^2}{\widehat{\theta}_N}.} \tag{30}$$

Proposition 5.1 (Sharp boundary for the specified exponential construction)**.** *Suppose $N \to \infty$, $m = N/K$ is even and $K/N^{7/8} \to c \in (0,\infty)$. For the half-block jackknife average and one Newton update,*

$$\sqrt{N}(\theta^{(1)} - \widehat{\theta}_N) \xrightarrow{p} -4\theta_0 c^4, \tag{31}$$

and hence

$$\sqrt{N}(\theta^{(1)} - \theta_0) \Rightarrow \mathcal{N}(-4\theta_0 c^4, \theta_0^2). \tag{32}$$

Proof. From Equation (28), the jackknife average has bias asymptotic to $-2\theta_0 m^{-2}$. Since $m = N/K \sim N^{1/8}/c$,

$$N^{1/4}(\mathbb{E}\theta^{(0)} - \theta_0) \to -2\theta_0 c^2. \tag{33}$$

The centered average fluctuates on scale $N^{-1/2}$, since its leading influence function averages all N observations and the remaining centered term has standard deviation $O((m\sqrt{K})^{-1}) = o(N^{-1/4})$. Likewise $\widehat{\theta}_N - \theta_0 = O_p(N^{-1/2})$. Therefore

$$N^{1/4}(\theta^{(0)} - \widehat{\theta}_N) \xrightarrow{p} -2\theta_0 c^2. \tag{34}$$

Use the exact identity Equation (30), $\widehat{\theta}_N \to_p \theta_0$, and the continuous mapping theorem to obtain Equation (31). The centralized MLE satisfies $\sqrt{N}(\widehat{\theta}_N - \theta_0) \Rightarrow \mathcal{N}(0, \theta_0^2)$, proving Equation (32). □

Remark 5.2 (Not a lower bound for all distributed estimators)**.** In the exponential model, each worker can transmit $\sum_{i\in\mathcal{D}_k} X_i$. Summing these K scalars recovers $\widehat{\theta}_N = N/\sum_i X_i$ *exactly*. Thus the sharp frontier in Theorem 5.1 concerns this particular jackknife-plus-one-step algorithm; it is not an information-theoretic limitation of distributed likelihood estimation.

6. Higher-order local extrapolation

6.1 Cancelling two bias terms with four quarter-block fits

Assume m is divisible by four. Let $A_{k,2}$ denote the average of MLEs on the two disjoint halves and $A_{k,4}$ the corresponding average on four disjoint quarter-blocks. Define

$$\boxed{J_k^{[3]} = \frac{8}{3}\widehat{\theta}_{k,m} - 2A_{k,2} + \frac{1}{3}A_{k,4}.} \tag{35}$$

The coefficients solve

$$\tfrac{8}{3} - 2 + \tfrac{1}{3} = 1, \tag{36}$$

$$\tfrac{8}{3} - 2(2) + \tfrac{1}{3}(4) = 0, \tag{37}$$

$$\tfrac{8}{3} - 2(2^2) + \tfrac{1}{3}(4^2) = 0. \tag{38}$$

The first identity preserves the leading influence term; the others cancel n^{-1} and n^{-2} terms of a common bias expansion. With a third-order bias expansion $\mathbb{E}R_n = b_1 n^{-1} + b_2 n^{-2} + b_3 n^{-3} + O(n^{-4})$, the local bias is

$$\mathbb{E}[J_k^{[3]}] - \theta_0 = 8b_3 m^{-3} + O(m^{-4}). \tag{39}$$

The analog of the L^2 triangle inequality establishes a centered remainder of order $O_p((m\sqrt{K})^{-1})$ after averaging workers, despite sharing nested data. Consequently

$$\theta^{(0,[3])} - \widehat{\theta}_N = O_p\left(m^{-3} + \frac{1}{m\sqrt{K}} + N^{-1}\right). \tag{40}$$

Corollary 6.1 (A sufficient 11/12 regime)**.** *Suppose Theorems 2.1 and 2.3 hold for $n \in \{m, m/2, m/4, N\}$, and the bias expansion through n^{-3} holds with $O(n^{-4})$ remainder. For m divisible by four and $K = o(N^{11/12})$, one global Newton step starting from $K^{-1}\sum_k J_k^{[3]}$ is first-order equivalent to centralized MLE.*

Proof. From Equation (40) and Theorem 4.1,

$$\sqrt{N}\|\theta^{(1,[3])} - \widehat{\theta}_N\| = O_p\left(\frac{K^6}{N^{11/2}} + \frac{K}{N^{3/2}} + N^{-3/2}\right) = o_p(1).$$

The first expression tends to zero if $K = o(N^{11/12})$; the others do so automatically. □

6.2 Exact higher-order bias in the exponential model

Using the exact mean in Equation (23), for $m > 4$,

$$\boxed{\mathbb{E}[J_k^{[3]}] - \theta_0 = \frac{8\theta_0}{(m-1)(m-2)(m-4)}.} \tag{41}$$

The bias magnitude is indeed $O(m^{-3})$. However, higher-order jackknife coefficients can amplify the finite-sample variance; the correction also requires local fits on quarter-blocks, which can have unstable or nonexistent moments for very small m. This is a reason to compare bias, variance, wall-clock time, and convergence failures rather than optimizing the asymptotic exponent alone.

6.3 A general bias–synchronization tradeoff

Consider a fixed bias-removal order $r \geq 1$. Suppose a local combination retains the same influence function and achieves

$$\theta_r^{(0)} - \widehat{\theta}_N = O_p\left(m^{-r} + \frac{1}{m\sqrt{K}} + N^{-1}\right), \tag{42}$$

with the requisite uniform moment bounds. After $q \geq 0$ locally regular global Newton updates, repeated quadratic error contraction yields

$$\|\theta_r^{(q)} - \widehat{\theta}_N\| = O_p\left[\left(m^{-r} + \frac{1}{m\sqrt{K}} + N^{-1}\right)^{2^q}\right]. \tag{43}$$

Proposition 6.2 (Sufficient rate envelope)**.** *For fixed integers $r \geq 1$ and $q \geq 0$, under Equation (42) and the required local quadratic contraction at each update, a sufficient condition for $\sqrt{N}(\theta_r^{(q)} - \widehat{\theta}_N) \to_p 0$ is*

$$\boxed{K = o\left(N^{\,1-1/(2^{q+1}r)}\right).} \tag{44}$$

Proof. With $m = N/K$, the bias component of the scaled remainder is $\sqrt{N}m^{-r2^q} = N^{1/2}(K/N)^{r2^q}$, which vanishes precisely under the stated sufficient regime. The centered-remainder component is $\sqrt{N}(\sqrt{K}/N)^{2^q}$. For $q = 0$ it equals $\sqrt{K/N}$, which vanishes since the displayed exponent is less than one; for $q \geq 1$ it is at most of order $N^{1/2-2^{q-1}}$ when $K \leq N$, hence vanishes. The full-sample term $\sqrt{N}(N^{-1})^{2^q}$ also vanishes. The sum of finitely many terms may be bounded by the corresponding sum of powers, finishing the proof. □

Table 1: Sufficient partition-growth exponents under fixed-dimensional regularity. These are *not* hardware speedups or universal lower bounds.

Initialization	Local bias order r	Global Newton steps q	$K = o(N^\alpha)$
Raw local mean	1	0	$\alpha = 1/2$
Raw local mean	1	1	$\alpha = 3/4$
Raw local mean	1	2	$\alpha = 7/8$
Half-block jackknife	2	0	$\alpha = 3/4$
Half-block jackknife	2	1	$\alpha = 7/8$
Quarter-block extrapolation	3	0	$\alpha = 5/6$
Quarter-block extrapolation	3	1	$\alpha = 11/12$

Remark 6.3. Theorem 6.2 is a rate-accounting result, conditional on an initializer satisfying Equation (42); it is *not* a proof that such an estimator exists for every r or every econometric model. Constructing stable higher-order initializers in heterogeneous or high-dimensional settings remains an open implementation and theory problem.

7. Numerical experiment: exact Gamma simulation

7.1 Design and reproducibility

To isolate estimator behavior from optimization and simulation discretization, we draw the sufficient statistics of exponential observations exactly. For each worker k, we generate four independent Gamma variates, each representing the sum of $m/4$ independent exponential observations; their sum is the block sufficient statistic. From these sums we evaluate local full, half, and quarter MLEs; the centralized MLE follows from the sum over all workers.

The common sample size is $N = 2^{20} = 1{,}048{,}576$, the true rate is $\theta_0 = 1$, and the partition sizes are $m \in \{256, 64, 16\}$, corresponding to $K \in \{4{,}096; 16{,}384; 65{,}536\}$. We run $B = 500$ independent Monte Carlo replications in each design with NumPy's default random generator initialized at seed `20261028`. These are *simulated data partitions, not physical computers.* We compare raw averaging with one or two Newton steps and both jackknife corrections with one Newton step. The estimand for each replication is the gap to the *same-sample centralized MLE*, scaled by $\sqrt{N}$:

$$\Delta_b = \sqrt{N}(\widehat{\theta}_b^{\mathrm{dist}} - \widehat{\theta}_{N,b}). \tag{45}$$

A gap to centralized MLE directly measures the numerical-statistical approximation, not the sampling error $\widehat{\theta}_{N,b} - \theta_0$. The exact code and the full 500-replication outputs are included in the accompanying reproducibility bundle.

Table 2: Mean scaled gap $\overline{\Delta}$ across 500 independent Monte Carlo replications. More negative numbers indicate a larger downward difference relative to centralized MLE. Rounding to six decimal places.

Estimator	$m = 256$	$m = 64$	$m = 16$
Raw mean + 1 Newton	-0.015728	-0.258367	-4.549613
Raw mean + 2 Newton	≈ 0	-0.000065	-0.020216
Half-block jackknife + 1 Newton	-0.000016	-0.000331	-0.093720
Quarter-block jackknife + 1 Newton	-0.000030	-0.000147	-0.011034

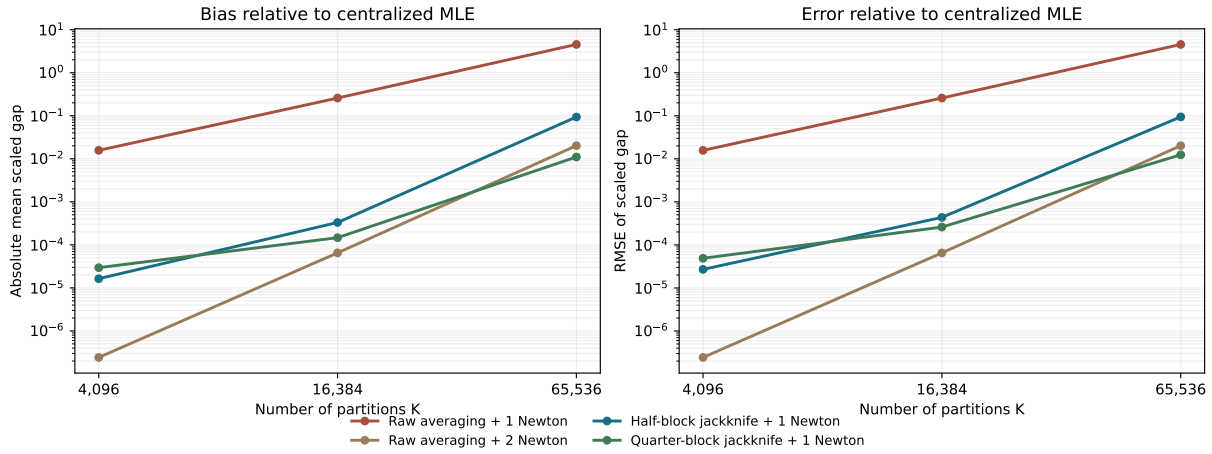


Figure 2: Absolute mean and root mean squared scaled gaps to the centralized MLE, calculated from the reproducibility script. Logarithmic axes make both small and large discrepancies visible. For the quarter-block method, its increased finite-sample variability can sometimes offset reduced bias.

7.2 Results

The half-block correction substantially reduces the one-step gap for large local samples, consistent with Theorem 4.2. At $m = 16$, higher-order correction improves the mean one-step gap still further, but also increases local fitting requirements and possible sensitivity to extreme quarter-sample MLEs. With two Newton steps, uncorrected averaging performs competitively: this is expected from Theorem 4.3, not evidence against the theory. The simulation is not a runtime evaluation and provides no evidence that one strategy is faster on actual distributed nodes.

7.3 What the simulation does not establish

Three caveats are material. First, N is fixed, whereas Theorems 4.2 and 6.1 describe triangular-array limits. Second, the Monte Carlo experiment uses exact aggregate statistics, hence omits network latency, local iterative optimizer cost, load imbalance, failure recovery, and floating-point differences across devices. Third, because the Gamma model has an elementary sufficient statistic, any empirical advantage here is a property of the estimator *approximation*, not an argument against exact aggregation.

8. Communication and computation accounting

8.1 Payload and synchronization rounds

For p parameters and dense symmetric Hessians, each worker's initial jackknife estimate is a p-vector. One subsequent derivative evaluation requires one p-vector and $p(p+1)/2$ unique Hessian entries per worker. Excluding protocol headers, numerical precision and broadcasts, the total amount of worker-to-coordinator data is on the order of

$$\boxed{K\left[2p + \frac{p(p+1)}{2}\right] \quad \text{numbers.}} \tag{46}$$

The two communication *phases* are not reducible to one if the second-phase score and Hessian are evaluated at the common, data-dependent $\theta^{(0)}$. One can transmit asynchronous or compressed derivatives, but then the current mathematical guarantee does not automatically apply. Dense-Hessian payload and coordinator matrix inversion may become prohibitive when p is large.

Table 3: Logical comparison. A "derivative phase" is one globally synchronized derivative evaluation after the initial communication of local estimates. Local-fit counts assume direct refitting on nested partitions.

Method	Local MLE fits / worker	Derivative phases	Sufficient exponent
Raw mean + 1 Newton	1	1	3/4
Raw mean + 2 Newton	1	2	7/8
Half-block jackknife + 1 Newton	3	1	7/8
Quarter-block extrapolation + 1 Newton	7	1	11/12

The meaningful systems-level claim is conditional: if an additional full-cluster synchronization has greater cost than two extra local likelihood fits, half-block jackknifing may be favorable; if local fits dominate and network latency is low, extra Newton steps may be faster. The cut-off depends on model dimension, optimizer complexity, scheduling, bandwidth, and numerical conditioning.

8.2 A cost model for an empirical follow-up

Let $T_{\mathrm{fit}}(m,p)$ denote the elapsed time for one local MLE fit, $T_{\mathrm{der}}(m,p)$ for one local derivative evaluation, L the per-phase synchronization latency, and B the effective worker-to-coordinator bandwidth. A first-order approximation is

$$T_{\mathrm{half+N}} \approx 3T_{\mathrm{fit}}(m,p) + T_{\mathrm{der}}(m,p) + 2L + \frac{bK[2p + p(p+1)/2]}{B}, \tag{47}$$

$$T_{\mathrm{raw+2N}} \approx T_{\mathrm{fit}}(m,p) + 2T_{\mathrm{der}}(m,p) + 3L + \frac{bK[3p + p(p+1)]}{B}, \tag{48}$$

where b is bytes per transmitted scalar and the expression deliberately suppresses overlapping parallelism, coordinator inversion, broadcasts, network contention, and data loading. These are accounting illustrations, *not* validated time-complexity theorems. An actual benchmark must record energy, cold starts, memory use, fit failures, p, hardware topology, and communication bytes independently of statistical accuracy.

9. Limitations and research agenda

9.1 Identification, misspecification, and dependence

The Fisher-information variance in the main theorem presumes a correctly specified regular model. Under misspecification, the asymptotic covariance is generally the sandwich form and the common influence function must be modified. If observations exhibit cross-node dependence or unequal distributions, the independent-block variance calculation in Theorem 3.1 changes. Fixed effects with incidental parameters can have nonstandard bias expansions; indeed, that setting is one of the motivations for Dhaene and Jochmans (2015). Neither validity nor exact exponents should be assumed to transfer without new proofs.

9.2 High-dimensional Hessians and nonconvex likelihoods

Our proof assumes fixed p. When p grows with N, the norm of the Fisher inverse and the constants in the high-moment expansion can depend on p, and the K exponent may weaken. In nonconvex likelihoods, local workers may converge to different stationary points, while a full Newton step can escape the basin in which the quadratic bound holds. Regularization, trust regions and generalized linear-system solvers may mitigate these issues but require separate theory. Stronger generality without control of such effects would overstate the current result.

9.3 Statistical versus computational optimality

We have not proved that the number of communications can be independent of N in a minimax sense, nor that this particular scheme achieves a lower bound over all algorithms. Indeed, exact sufficient-statistic aggregation defeats any supposed general lower bound in the exponential model. The appropriate open question is narrower: *over a specified class of nonlinear likelihoods, and under a model of allowable local computation and communication, what Pareto frontier links local refits, derivative synchronizations, transferred bytes, and statistical precision?*

A useful research program would (i) characterize the finite-sample risk of high-order jackknife starts; (ii) derive feasible model-specific analytical bias corrections that use fewer local optimizations; (iii) compare with modern multi-round techniques such as CEASE (Fan et al., 2023); (iv) extend the expansions to clustered or time-dependent panels; and (v) implement a wall-clock benchmark across physical nodes, varying the cost ratio of fitting to synchronization.

10. Conclusion

This note develops and verifies a local-bias-correction-plus-global-Newton construction for distributed maximum likelihood estimation. Under explicit uniform mean-square remainders, second-order bias expansions, and local Hessian regularity, a half-block jackknife initializer followed by one global Newton step is first-order equivalent to centralized MLE when $K = o(N^{7/8})$. The stochastic component of the jackknife remainder has been kept throughout the proof. A third-order local extrapolation yields the sufficient exponent $11/12$ under stronger assumptions. Exact Gamma calculations establish the rate and a nonzero boundary shift in the exponential model, while a reproducible Monte Carlo experiment confirms the expected qualitative pattern.

The scientific interpretation is appropriately limited. Related distributed one-step estimation and

higher-order jackknife ideas are established, and the method has not yet been benchmarked in a real networked environment. The potentially productive direction is the *quantitative exchange between local correction and global synchronization*, rather than an unsupported claim to have surpassed all existing maximum likelihood algorithms.

Statements and reproducibility

Code and data. The accompanying package contains `monte_carlo.py`, the simulation CSV outputs, and figure sources. The simulation requires Python, NumPy, and Matplotlib. All generated observations are synthetic; no proprietary data or user data were used.
Authorship. The designation "GPT-6.1" is included at the request of the user preparing this draft. It refers to an AI-assistance authorship label, rather than a verified human contributor. Responsibility for scientific validation and compliance with any target journal's authorship policy rests with the submitting researcher.
Research status. Unsubmitted methodological working paper. Theoretical claims are explicitly assumption-dependent. Priority and novelty have not been exhaustively established.

References

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363. https://www.jmlr.org/papers/v14/zhang13b.html.

Huang, C. and Huo, X. (2019). A distributed one-step estimator. *Mathematical Programming*, 174(1–2):41–76. https://doi.org/10.1007/s10107-019-01369-0.

Dhaene, G. and Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3):991–1030. https://doi.org/10.1093/restud/rdv007.

Fan, J., Guo, Y., and Wang, K. (2023). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, 118(542):1000–1010. https://doi.org/10.1080/01621459.2021.1969238.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia. https://doi.org/10.1137/1.9781611970319.

A. Proof details for the exponential remainder

Let $Y = \theta_0 \bar{X}_n$, so $nY \sim \text{Gamma}(n, \text{rate} = 1)$. We have $\widehat{\theta}_n = \theta_0/Y$ and $R_n = \theta_0(1-Y)^2/Y = \theta_0(Y^{-1} - 2 + Y)$. The requisite moments for $n > 2$ are

$$\mathbb{E}Y = 1, \quad \mathbb{E}Y^2 = 1 + 1/n, \quad \mathbb{E}Y^{-1} = \frac{n}{n-1}, \quad \mathbb{E}Y^{-2} = \frac{n^2}{(n-1)(n-2)}. \tag{49}$$

Therefore

$$\begin{aligned}
\mathbb{E}R_n &= \theta_0\left(\frac{n}{n-1} - 2 + 1\right) = \frac{\theta_0}{n-1},\\
\mathbb{E}R_n^2 &= \theta_0^2\mathbb{E}(Y^{-2} - 4Y^{-1} + 6 - 4Y + Y^2)\\
&= \theta_0^2\left[\frac{n^2}{(n-1)(n-2)} - \frac{4n}{n-1} + 3 + \frac{1}{n}\right]\\
&= \theta_0^2\frac{3n+2}{n(n-1)(n-2)}.
\end{aligned}$$

This supplies an elementary verification of Equation (27). For nested subsamples, the same ψ is used for each sample, which is why the linear term cancels in the jackknife identity.

B. Exact bias of the quarter-block extrapolation

At full, half and quarter-block sizes, the rate MLE has expectations $\theta_0 m/(m-1)$, $\theta_0 m/(m-2)$ and $\theta_0 m/(m-4)$ respectively. Inserting these values into Equation (35) yields

$$\begin{aligned}
\mathbb{E}J_k^{[3]} - \theta_0 &= \theta_0\left[\frac{8/3}{m-1} - \frac{4}{m-2} + \frac{4/3}{m-4}\right]\\
&= \frac{8\theta_0}{(m-1)(m-2)(m-4)},
\end{aligned}$$

proving Equation (41). The high-order bias coefficients are positive in this particular model, whereas the half-block bias in Equation (28) is negative. Such alternating signs are one reason direct extrapolation is not necessarily monotone in finite-sample accuracy.