

Tail-Accuracy Degradation and Optimal Block Length for Block Self-Normalized Quantile Empirical Saddlepoint Inference under Dependence

Yucheng Tao

YuchengTao2000@126.com

Guangxi Normal University <https://orcid.org/0009-0009-6644-9779>

Research Article

Keywords: self-normalization, empirical saddlepoint, quantile inference, α -mixing, optimal block length

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11352599/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Tail-Accuracy Degradation and Optimal Block Length for Block Self-Normalized Quantile Empirical Saddlepoint Inference under Dependence

Yucheng Tao ^{1,*} 

¹ School of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, China; YuchengTao2000@126.com

Abstract

Self-normalization combined with empirical saddlepoint approximation reaches higher-order tail accuracy under independent sampling, but under dependence the degradation is unquantified and the block length lacks a basis. This paper quantifies both for the block self-normalized quantile pivot under strong (α)-mixing. Self-normalizing the block sums by their own quadratic form frees the pivot from the marginal distribution, and expanding its cumulants by multilinearity places the tail at the $n^{-1/2}$ rather than the $m^{-1/2}$ scale. The relative tail error obeys an explicit law, $\text{err}(\ell) = -C_1^{\text{eff}} n^{-1/2} + D(\ell) + O(n^{-1})$, led by a previously unnoticed variance-inflation floor $E_\infty(\ell) \sim c_\Phi / \ell$ that does not vanish with the number of blocks but converges to a positive limit. Balancing that floor against a skewness transient of the same order gives the optimal block length $\ell^* \propto \sqrt{n}$ in closed form, with an exponent independent of the dependence; no interior optimum exists when $\lambda \leq 0$ or $\tau \geq 1/2$. The lattice structure of the pivot also forces any continuous tail approximation to err by at least $\Theta(m^{-1}\ell^{-3/2})$, a bound that is attained. The two main results combine into an interval whose coverage error is $O(n^{-1})$ near ℓ^* , turning the block length into a computable quantity; coverage is $95\% \pm 0.15\%$ over 16 configurations, against 60.1% for the first-order baseline.

Keywords: self-normalization; empirical saddlepoint; quantile inference; α -mixing; optimal block length

MSC: 62G20; 62E20; 62G09; 60F05

1. Introduction

Quantile inference underlies the standard risk measures, and it is constrained by a structural difficulty: the asymptotic variance of the sample quantile is proportional to the reciprocal of the density at the quantile, $1/f(q_\tau)$, and estimating this reciprocal stably is hard in small samples, in heavy tails, and at extreme quantiles. Interval procedures built on a kernel estimate of that density, such as the Wald and Hall–Sheather intervals [1], are therefore sensitive to bandwidth selection and show noticeable coverage fluctuations in finite samples.

Frequentist procedures for quantiles under dependence fall into two families, and each buys its validity with a sacrifice. The first allows dependence but gives up either higher-order accuracy or density-freeness. HOGA and SCHULZ [2] use self-normalization for time-series quantile and expected shortfall regressions, but their self-normalizer is built

Received:

Accepted:

Published:

Copyright: © 2026 by the author.

Submitted to *Mathematics* for possible open access publication under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

from an expanding window, their results are first-order only, and their assumptions require a conditional density that exists and is Lipschitz near the origin, so the procedure is not density-free. HOUNYO and LIN [3] treat two-way clustered quantile regression with a sandwich-type cluster-robust variance rather than self-normalization; their bread is a kernel density estimate, the results are again first-order Gaussian, and the authors acknowledge that higher-order bandwidth optimization lies beyond their scope. EINMAHL and HE [4] achieve density-freeness through a Galois connection, but only for independent heterogeneous data. The second family attains higher-order accuracy together with density-freeness, by combining self-normalization, which replaces an explicit scale estimate by a quadratic form internal to the statistic, with an empirical saddlepoint approximation of the tail probability; under independent sampling the combination reaches $O(n^{-3/2})$ tail accuracy. SNQESA [5] belongs to this family: it self-normalizes the indicator score into a pivot and evaluates the tail probability by a constrained saddlepoint approximation, without estimating $f(q_\tau)$, but its theorems and simulations are all carried out under independent sampling, and the mixing case appears in its discussion only as a feasible extension. The two families have not been joined, and this paper occupies exactly their junction: SNQESA supplies the engine for density-free higher-order accuracy on the quantile side and LAHIRI's block framework the expansion tools on the dependence side, and a degradation law with an optimal block length can appear on the quantile side only once the two are combined. For a fixed quantile level $\tau \in (0, 1)$, write $g_i(t) = \mathbf{1}\{X_i \leq t\} - \tau$ for the indicator score and let $\bar{g}_j(t)$ be its average over the j -th block of length ℓ . We study the block self-normalized quantile pivot

$$T_n(t) = \frac{\sum_{j=1}^m \bar{g}_j(t)}{\sqrt{\sum_{j=1}^m \bar{g}_j(t)^2}}, \quad m = \lfloor n/\ell \rfloor,$$

computed from a strongly (α)-mixing sample, with λ the lag-1 correlation of the indicator process.

Two questions follow about this pivot, and neither has an answer once the sampling is dependent.

(Q1) How much does the tail accuracy of T_n degrade, relative to the Gaussian benchmark?

(Q2) How should the block length ℓ be chosen in response?

The first is quantitative: the Gaussian benchmark against which a self-normalized pivot is usually read carries variance 1, whereas the block pivot carries a long-run variance that dependence inflates. The second is a tuning problem which, in the quantile case, has no existing answer.

What is already known about the degradation is confined to sums and means. LAHIRI [6] gives three side-by-side error sources for studentized means and M -estimators, arising from $n^{-1/2}$, from the number of blocks through $m^{-1/2}$, and from studentizing bias; LAHIRI [7] gives an Edgeworth expansion for sums of block variables in which the error is carried by two series, in $n^{-1/2}$ and in $[n/\ell]^{-1/2}$; WANG and POLONIK [8] show that for smooth functionals of the block empirical likelihood the coverage error drops from $O(m^{-1})$ to $O(m^{-2})$. On the quantile side the closest work is LAHIRI and SUN [9], which has only a first-order Berry–Esseen bound. On the block-method side, KUFFNER et al. [10] give an optimal block length for the quantile block bootstrap, but that optimality is in the sense of bootstrap convergence rates; it is neither an analytic relative error nor self-normalized. Table 1 places these boundaries side by side. Dependence, quantiles and self-normalization are pairwise studied, yet their intersection is empty: on the quantile side not even the analytic form of the degradation or a closed form for the optimal block length is available.

Table 1. Work directly related to this paper and the boundaries it leaves open.

Work	Dependence	Self-normalization	Density-free	Accuracy
SNQESA [5]	Independent	Scalar $S_n / \sqrt{Q_n}$	Yes	Tail $O(n^{-3/2})$
Hoga–Schulz [2]	α -mixing	Expanding window	No	First-order CLT
Hounyo–Lin [3]	Two-way clustering	No (sandwich)	No	First-order Gaussian
Einmahl–He [4]	Independent, heterog.	Rank-type variance	Yes	First-order normal
Lahiri [6]	Weak dependence	N/A	N/A	Three-column structure
Lahiri–Sun [9]	Polynomial mixing	N/A	N/A	First-order Berry–Esseen
This paper	α -mixing	Block self-normalization	Yes	See Theorem 2

Dependence changes the picture entirely, and it is why neither family settles the question. The variance of the score sum is no longer $\tau(1 - \tau)$ but a long-run variance, and the algebraic identity on which density-freeness rests no longer holds, so the two routes lose their support at the same time. One further point has gone unnoticed: the block pivot takes lattice values, so its tail function is a step function rather than a smooth one, and this imposes a lower bound on the attainable uniform accuracy.

Our starting point is a pair of neglected quantities. Measuring the tail error against the Gaussian benchmark, $\text{err}(\ell) = \mathbb{P}(T_n \geq x) / (1 - \Phi(x)) - 1$, at a fixed block length it does not vanish as the number of blocks grows; it converges to the positive limit $E_\infty(\ell) \sim c_\Phi / \ell$, with

$$c_\Phi = \frac{x\phi(x)\lambda}{(1 - \Phi(x))(1 - \lambda^2)},$$

a variance-inflation floor that arises because the long-run variance of the block pivot is inflated by dependence while the benchmark assumes variance 1. The transient on the other side comes from bias and skewness. The bias term $\frac{m_1}{\sigma_T}\phi(x)$ is of order $q_1\phi(x)/(2\sqrt{n})$, and because $\text{pred}(\lambda, \ell) \sim p_1/\sqrt{\ell}$ with $p_1 = -2q_1$, the skewness term is likewise independent of ℓ ; together they give a transient that does not depend on ℓ . Balancing the floor against this transient gives an interior optimum in ℓ .

Both questions have quantitative answers, and they are given by four results. The first removes a nuisance: the pivot is density-free, in that the distribution of $T_n(t)$ depends only on the indicator process and not on F , so no density estimate and no bandwidth beyond the block length itself enter what follows; a truncation construction shows that the setting is nonempty. The second answers (Q1), by resolving the relative tail error into an explicit two-term expansion,

$$\text{err}(\ell) = -\frac{C_1^{\text{eff}}(x, \tau)}{\sqrt{n}} + D(\ell) + O(n^{-1}), \quad C_1^{\text{eff}}(x, \tau) = \frac{q_1(2x^2 + 1)\phi(x)}{6(1 - \Phi(x))},$$

in which the first term collects the $n^{-1/2}$ -order bias and skewness contributions and the second is that same floor, written $D(\ell)$, which takes one of three forms according to the decay of the mixing coefficients. The third answers (Q2), since balancing the two terms gives the interior-optimal block length

$$\ell^* = \frac{6x\lambda}{(1 - \lambda^2)(2x^2 + 1)q_1} \sqrt{n} (1 + o(1)), \quad q_1 = \frac{(1 - 2\tau)(1 + 4\lambda + \lambda^2)}{\sqrt{\tau(1 - \tau)(1 - \lambda)(1 + \lambda)^3}},$$

proportional to $\sqrt{n}$ with an exponent independent of λ . The fourth is separate from the two questions: the lattice structure of the pivot forces the uniform error of any continuous approximation of the tail to be at least of order $\Theta(m^{-1}\ell^{-3/2})$, and a piecewise-linear construction attains it. Together the first three yield a usable procedure, which is the form in which the answers are applied: inverting the tail probability gives a quantile interval whose coverage error is $O(n^{-1})$ when ℓ is chosen near ℓ^* , and reverts to $O(n^{-1/2})$

otherwise. Numerically, the method attains coverage within $95\% \pm 0.15\%$ across 16 (λ, ℓ) configurations, whereas the first-order Gaussian baseline attains only 60.1% at $\lambda = 0.8$ and $\ell = 2$; the closed-form interval length is within 2% of the exact value, and the computational cost is 0.0037 milliseconds per replication.

The rest of the paper is organized as follows. Section 2 (Materials and Methods) sets out the problem, the block self-normalized pivot and the regularity assumptions, proves density-freeness and gives the truncation construction. Section 3 (Results) collects the four results: the tail-accuracy degradation law and the variance-inflation floor in Section 3.1, the scaling law and leading-order constant for the optimal block length in Section 3.2, the uniform-accuracy lower bound induced by the lattice structure and its attainability in Section 3.3, and the simulation study in Section 3.4. Section 4 discusses the questions that remain open and the limits of applicability, and Section 5 concludes. All proofs and derivations are collected in Appendix A.

2. Materials and Methods

This section introduces the notation and the basic objects, and explains why the pivot is independent of F at the true quantile. Let $X_1, \dots, X_n$ come from a strictly stationary strong (α) -mixing process whose mixing coefficients satisfy $\alpha(k) \leq Ck^{-\beta}$ with $\beta > 1$ (for the definition see [11]). Fix $\tau \in (0, 1)$, write $q_\tau = F^{-1}(\tau)$, and define the indicator score

$$g_i(t) = \mathbf{1}\{X_i \leq t\} - \tau, \quad \mathbb{E}g_i(q_\tau) = 0.$$

Take a block length ℓ that diverges with n while satisfying $\ell/n \rightarrow 0$, and let the number of blocks be $m = \lfloor n/\ell \rfloor$. Write the j -th block mean as $\bar{g}_j(t) = \frac{1}{\ell} \sum_{i \in B_j} g_i(t)$; the block self-normalized pivot is

$$T_n(t) = \frac{\sum_{j=1}^m \bar{g}_j(t)}{\sqrt{\sum_{j=1}^m \bar{g}_j(t)^2}}. \quad (1)$$

The goal of inference is to invert the tail probability of T_n into a confidence interval for q_τ . The scale is the within-block quadratic form $\sum_j \bar{g}_j^2$ rather than a lag-covariance estimate of the long-run variance as in LAHIRI [6]: the former is independent of F at $t = q_\tau$ (Theorem 1), and it requires no separate bandwidth for the variance.

Assumption 1. (A1) The process is strictly stationary with $\alpha(k) \leq Ck^{-\beta}$ and $\beta > 1$. (A2) F is continuous in a neighbourhood of q_τ and $F(q_\tau) = \tau$. (A3) $x \in [a, b] \subset (0, \infty)$. (A4) The cumulants of the pivot satisfy, uniformly on $[a, b]$, $\kappa_2(T_n) = S_{11} + O(m^{-1}\ell^{-1})$, $\kappa_3(T_n) = \frac{\Lambda^{111} - 3S_{11}S_{12}}{\sqrt{m}} + O(n^{-1})$ and $\kappa_4(T_n) = O(n^{-1})$, and the resulting Edgeworth expansion has remainder $O(n^{-1})$; the block sums $\{B_t\}$ satisfy a Bernstein-type deviation inequality.

(A4) is stronger than (A1); the trade-offs among its three regularities and the sufficient condition for each are given in the two remarks that open Appendix A.2.

Theorem 1. The pivot $T_n(t)$ is a function of the indicator process $\{\mathbf{1}\{X_i \leq t\}\}$, so its distribution depends only on the joint law of that process and not on F . The constants in the degradation law, C_1^{eff} , C_κ and ℓ^* , are therefore also independent of F .

This conclusion is not vacuous. Given an arbitrary stationary binary sequence $\{Z_i\}$ and independent uniform random variables U_i , set

$$X_i = \begin{cases} F^{-1}(U_i\tau), & Z_i = 1, \\ F^{-1}(\tau + U_i(1 - \tau)), & Z_i = 0. \end{cases}$$

Then X_i has marginal distribution exactly F , and at the true quantile $\mathbf{1}\{X_i \leq q_\tau\} = Z_i$. Hence any F and any dependence structure of the indicator process can be realized at $t = q_\tau$, and the derivation of the degradation law needs to handle only the indicator process, not the specific form of F . The scope of this statement should be noted: the identity uses $F(q_\tau) = \tau$ and therefore holds only at $t = q_\tau$. The whole path of $T_n(t)$ in a neighbourhood of q_τ , which is what interval inversion requires, still depends on F ; the density-free assertion of this paper applies to the distribution of the pivot at $t = q_\tau$ alone.

3. Results

3.1. Tail-Accuracy Degradation Law

For a two-state Markov indicator process, the long-run variance and the third cumulant of the block sums are available in closed form. Let λ denote the lag-1 correlation of the indicator process, set $\sigma^2 = \tau(1 - \tau)$, and let $\lambda_3 = \kappa_3(T_n)/\kappa_2(T_n)^{3/2}$ be the standardized skewness. Then

$$\sqrt{m} \lambda_3(T_n) = \frac{\Lambda^{111} - 3S_{11}S_{12}}{S_{11}^{3/2}} + O(m^{-1}), \quad S_{11} = 1 + \frac{2\lambda(1 - \lambda^\ell)}{(1 - \lambda)^2 R_2(\ell)}, \quad (2)$$

where $R_2(\ell) = \ell + 2 \sum_{d \geq 1} (\ell - d) \lambda^d$. These three quantities enter only through the second and third cumulants of the indicator process, so they hold for any stationary binary process. Block averaging suppresses the skewness of the block sums by a factor $\sqrt{\ell}$, that is, $S_{12}, \Lambda^{111} = O(\ell^{-1/2})$, whence $\kappa_3(T_n) = O(n^{-1/2})$ and $\kappa_4(T_n) = O(n^{-1})$ (Lemma A4). The natural expansion scale of the degradation law is therefore $n^{-1/2}$, not the $m^{-1/2}$ suggested by the number of blocks.

Theorem 2. Under Assumption 1,

$$\frac{\mathbb{P}(T_n(t) \geq x)}{1 - \Phi(x)} - 1 = -\frac{C_1^{\text{eff}}(x, \tau)}{\sqrt{n}} + D(\ell) + O(n^{-1}),$$

where $D(\ell)$ is the variance-inflation floor; under geometric mixing $D(\ell) \sim C_\kappa/\ell$ with $C_\kappa = \lambda x \phi(x)/[(1 - \lambda^2)\Phi(-x)]$. Here $C_1^{\text{eff}}(x, \tau) = \frac{q_1(2x^2+1)\phi(x)}{6(1-\Phi(x))}$ collects all $n^{-1/2}$ -order skewness and bias contributions, with q_1 given in Lemma 2. Under general mixing, $D(\ell)$ splits into three branches according to the decay rate $\alpha(k) \leq Ck^{-\beta}$,

$$D(\ell) \sim \begin{cases} C/\ell, & \beta > 2, \\ C \ell^{-1} \log \ell, & \beta = 2, \\ C \ell^{1-\beta}, & 1 < \beta < 2. \end{cases}$$

Equivalently, $\text{err}(\ell) = |-C_1^{\text{eff}}/\sqrt{n} + D(\ell)|(1 + o(1))$. When $\lambda > 0$ we have $C_\kappa > 0$ and the floor is positive; when $\lambda \leq 0$ we have $C_\kappa < 0$ and the floor becomes a ceiling (see Remark 1).

The scale variable of the degradation law is the sample size n , not the number of blocks m : the leading term comes from standardized skewness and bias, is of order $n^{-1/2}$, has a coefficient proportional to $1 - 2\tau$, and vanishes at $\tau = 1/2$. The floor $D(\ell)$, which depends on ℓ , gives the second layer of the structure. This two-layer structure for sums and means was established by LAHIRI [6]; the contribution of this paper is to carry it over to quantile inference, indicator scores and self-normalization. The derivation of Theorem 2 is given in Appendix A.3.

3.2. Optimal Block Length

The second term $D(\ell)$ of Theorem 2 varies with ℓ and, in earlier analyses, was treated as a negligible higher-order term. With the Gaussian as benchmark, the leading term of the error in the ℓ direction is the variance-inflation floor

$$E_{\infty}(\ell) = \frac{1 - \Phi(x/\sqrt{S_{11}(\ell)})}{1 - \Phi(x)} - 1 \sim \frac{c_{\Phi}}{\ell}, \quad c_{\Phi} = \frac{x\phi(x)\lambda}{(1 - \Phi(x))(1 - \lambda^2)}, \quad (3)$$

which decreases as ℓ grows but does not vanish as the number of blocks grows. The next lemma writes this limit in explicit form.

Lemma 1. *For fixed ℓ , as $m \rightarrow \infty$ the error $\text{err}(\ell)$ converges to $E_{\infty}(\ell)$, and as $\ell \rightarrow \infty$ we have $E_{\infty}(\ell) \sim c_{\Phi}/\ell$. When $\lambda > 0$, $c_{\Phi} > 0$ and the limit is positive.*

Remark 1. *The sign of the floor is determined by the sign of λ . Since $S_{11}(\ell) - 1 \sim \frac{2\lambda}{1-\lambda^2} \cdot \frac{1}{\ell}$, when $\lambda < 0$ we have $S_{11}(\ell) < 1$: the long-run variance of the block pivot is smaller than 1, $E_{\infty}(\ell) < 0$, and the floor degenerates into a ceiling. The sign of the transient changes accordingly, the two terms no longer cancel, and the closed form for ℓ^* does not apply when $\lambda \leq 0$. The optimal-block-length results of this paper are therefore confined to $\lambda > 0$.*

The transient on the other side comes from bias and skewness. The large- ℓ limit of (2) gives $\text{pred}(\lambda, \ell) \sim p_1/\sqrt{\ell}$ and $S_{12}(\lambda, \ell) \sim q_1/\sqrt{\ell}$ with $p_1 = -2q_1$, so the transient does not depend on ℓ .

Lemma 2. *The transient due to bias and skewness satisfies*

$$\text{tr}(\ell, n) = \frac{m_1}{\sigma_T} \phi(x) + \frac{\lambda_3}{6} (x^2 - 1) \phi(x) \sim -\frac{q_1(2x^2 + 1)\phi(x)}{6\sqrt{n}},$$

where

$$q_1 = \frac{(1 - 2\tau)(1 + 4\lambda + \lambda^2)}{\sqrt{\tau(1 - \tau)}(1 - \lambda)(1 + \lambda)^3},$$

which does not involve ℓ .

The floor decreases with ℓ while the transient does not depend on ℓ , and the point where the two are equal is the interior optimum for the block length.

Theorem 3. *Under the conventions of Theorem 2 (that is, $\lambda > 0$ and $\tau < 1/2$), the interior-optimal block length for tail accuracy is*

$$\ell^* = \frac{6x\lambda}{(1 - \lambda^2)(2x^2 + 1)q_1} \sqrt{n} (1 + o(1)). \quad (4)$$

Proofs of Lemma 1, Lemma 2 and Theorem 3 are given in Appendix A.4. The constant has three directly checkable properties. It is proportional to $\lambda/(1 - \lambda^2)$, so stronger dependence shifts the block length to the right. It is proportional to $1/(1 - 2\tau)$, so as $\tau \rightarrow 1/2$ the transient vanishes, the floor has nothing to cancel against, and no interior optimum exists; for $\tau > 1/2$ we have $q_1 < 0$ and the closed form returns a negative block length, which shows that this range likewise falls outside the analysis. The exponent of n does not depend on λ , which enters only the constant. Figure 1 sketches the mechanism.

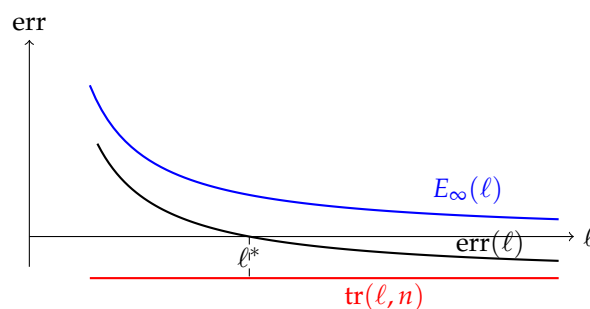


Figure 1. The variance-inflation floor $E_\infty(\ell)$ decreases with ℓ while the transient $\text{tr}(\ell, n)$ does not depend on ℓ ; the two curves cross at ℓ^* , hence $\ell^* \propto \sqrt{n}$.

Inverting the tail probability yields a quantile interval. For each t , the two-sided p -value for testing $H_0 : F(t) = \tau$ is

$$p(t) = 2 \min\{\mathbb{P}_0(T_n(t) \geq T_n^{\text{obs}}(t)), \mathbb{P}_0(T_n(t) \leq T_n^{\text{obs}}(t))\}, \quad (227)$$

where $\mathbb{P}_0$ denotes the distribution of the pivot under H_0 , which is independent of F by Theorem 1, and $T_n^{\text{obs}}(t)$ is computed from the sample. The confidence interval of level $1 - \alpha$ is $\hat{I}_\alpha = \{t : p(t) \geq \alpha\}$, with endpoints obtained from $p(t) = \alpha$.

Its coverage order follows from the degradation law. At $\ell \asymp \sqrt{n}$ the relative error of the tail probability is $|-C_1^{\text{eff}}/\sqrt{n} + D(\ell)|$, and the two terms have opposite signs (for $\tau < 1/2$ we have $C_1^{\text{eff}} > 0$, so the first term is negative, while $D(\ell) > 0$ for $\lambda > 0$). Taking $\ell = \ell^*$ makes them cancel exactly, the error drops to $O(n^{-1})$, and the coverage error of $\hat{I}_\alpha$ is therefore $O(n^{-1})$, that is, second-order coverage. When the block length departs from ℓ^* the two terms no longer cancel and the coverage error reverts to $O(n^{-1/2})$; this is precisely the role of the block length. Numerically (analytic tail, $\lambda = 0.5$, $x = 2$), at $\ell = \ell^*(n)$ we obtain $n \text{ err} = 16$ to 31 for $n = 2^{10}, \dots, 2^{30}$, a bounded quantity, consistent with $O(n^{-1})$.

3.3. Uniform-Accuracy Lower Bound Induced by the Lattice Structure

The block pivot takes lattice values. The pair $(S_{ii}, Q_{ii,jj})$ lives on an integer lattice and $T_{ii,jj} = S_{ii}/\sqrt{Q_{ii,jj}}$, so the tail function $P_{\text{ex}}(x) = \mathbb{P}(T_n \geq x)$ is an exact step function whose jumps occur at $x_{ii,jj} = S_{ii}/\sqrt{Q_{ii,jj}}$, with jump size equal to the sum of dp over all lattice points falling at that x . Different pairs (ii, jj) can give the same x ; in that case the jump size is their sum and need not equal a single $dp[ii, jj]$.

Theorem 4. Let $\mathcal{B}[a, b]$ be the set of lattice points whose jumps fall in $[a, b]$, let $w(x)$ be the jump size at a jump point x , and let $w_{\max} = \max_{x \in [a, b]} w(x)$. Then for any continuous function h ,

$$\sup_{x \in [a, b]} |P_{\text{ex}}(x) - h(x)| \geq \frac{1}{2} w_{\max}. \quad (247)$$

Under a local central limit theorem $w_{\max} = \Theta(m^{-1}\ell^{-3/2})$, and therefore

$$\inf_{h \in C^0} \sup_{x \in [a, b]} |P_{\text{ex}}(x) - h(x)| = \Theta(m^{-1}\ell^{-3/2}). \quad (249)$$

The proof of Theorem 4 is given in Appendix A.5. This lower bound shows that continuous corrections such as the area-fraction correction or the Euler–Maclaurin correction cannot push the uniform accuracy below $m^{-1}\ell^{-3/2}$. Its relation to Theorem 3 should be stated: what is bounded here is the uniform absolute approximation error of the tail function by a continuous function, whereas $\text{err}(\ell)$ is a fixed-point relative error. The two

Table 2. Coverage (%) and average interval length, $\lambda = 0.6$, $\tau = 0.3$; lengths use $B = 2000$ replications and coverage uses $B = 4 \times 10^4$. L_{or} is the ratio of the ESA-an length to the oracle length.

ℓ	m	ESA-an	SN-N	MBB-pct	MBB-t	BEL	L_{ESAan}	L_{or}
4	16	95.31	86.56	81.40	84.20	81.30	.3513	1.011
4	64	95.00	86.89	83.55	84.20	83.65	.1887	1.001
8	64	94.86	91.24	90.40	91.00	88.65	.1387	1.002
16	64	95.05	93.47	91.45	92.40	89.65	.1011	1.002

conventions differ, and the lower bound does not directly constrain the quantity being optimized. At the interior optimum $\ell^* \asymp \sqrt{n}$ we have $m^{-1}\ell^{-3/2} \asymp n^{-5/4}$, far below $n^{-1/2}$, so it is not binding there either. What it characterizes is the intrinsic limitation of continuous approximations at a given ℓ .

In the opposite direction, the contribution of the lattice structure to the fixed-point tail error is controlled by the same order.

Corollary 1. *Under the local limit conditions of Theorem 4, the difference between the continuous Edgeworth tail Φ_{EE} and the exact step function satisfies*

$$\sup_{x \in [a,b]} |P_{\text{ex}}(x) - \Phi_{\text{EE}}(x)| \leq 2w_{\text{max}} + O(n^{-1}) = O(n^{-1}),$$

where the lattice term satisfies $2w_{\text{max}} = O(m^{-1}\ell^{-3/2})$, which at the interior optimum $\ell^* \asymp \sqrt{n}$ is $O(n^{-5/4}) = o(n^{-1})$. The lattice structure therefore contributes only an oscillation term below the remainder of the continuous part; the leading term remains the $O(n^{-1})$ of (A4), and the lattice case introduces no additional assumption.

3.4. Simulation Study

The experiments take $F = U(0,1)$ together with several general F , with a two-state Markov chain as the indicator process and $\tau = 0.3$; all computations are carried out on a GPU. The coverage curves fix the number of blocks at $m = 64$ and vary the block length $\ell \in \{2, 4, 8\}$ and the dependence strength $\lambda \in [0, 0.8]$, corresponding to $n = 128, 256, 512$. The joint comparison of coverage and interval length is carried out at $\lambda = 0.6$, with $B = 2000$ replications for lengths and $B = 4 \times 10^4$ for coverage; computational cost is benchmarked at $n = 1024$. Five methods are compared: an oracle given by the exact FFT joint law, the procedure of this paper (ESA-an), which uses the closed-form Edgeworth tail, the first-order normal-critical-value method (SN-N), the percentile and studentized versions of the moving block bootstrap [12,13], and the block empirical likelihood [14].

Figure 2 reports exact coverage, computed directly from the FFT joint law and therefore free of Monte Carlo error. The curve for ESA-an hugs 95% for all $\lambda \in [0, 0.8]$ and $\ell \in \{2, 4, 8\}$, with a range within $\pm 0.15\%$. The first-order Gaussian baseline deteriorates rapidly as dependence strengthens, and the smaller the block length the worse it becomes: at $\lambda = 0.8$, $\ell = 8$ still holds 84.7%, $\ell = 4$ falls to 74.5%, and $\ell = 2$ is down to 60.1%.

Table 2 compares coverage and average interval length at $\lambda = 0.6$. The interval length of ESA-an is within 2% of the exact oracle. All three baselines substantially undercover: SN-N lies between 86.6% and 93.5%, the percentile version of the moving block bootstrap between 81.4% and 91.5%, its studentized version slightly better at 84.2% to 92.4%, and the block empirical likelihood between 81.3% and 89.7%. These baselines in fact produce shorter intervals; at $\ell = 16$, for instance, the studentized bootstrap gives 0.0927 against 0.1011 for ESA-an.

On computational cost, the critical value of ESA-an comes from a single closed-form evaluation and the interval inversion is an integer bisection, so each replication takes about

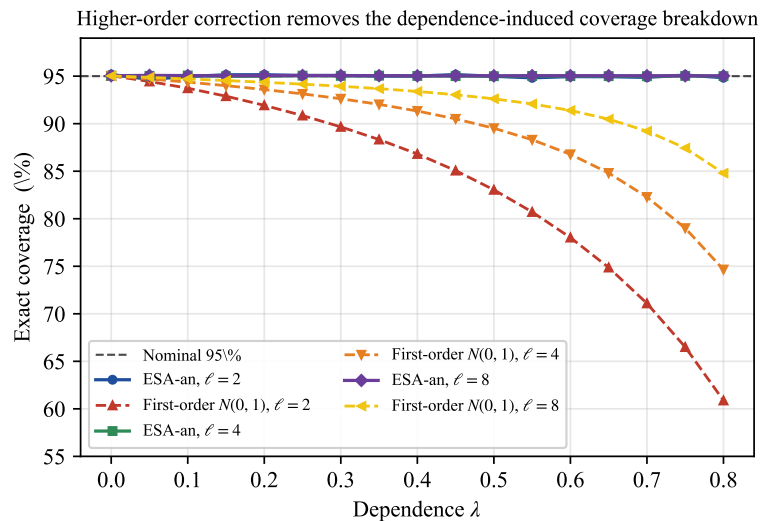


Figure 2. Coverage as a function of the dependence strength λ (exact values, $\tau = 0.3$, $m = 64$). ESA-an hugs 95% for every λ and ℓ ; the first-order $N(0, 1)$ baseline deteriorates as λ grows, and the smaller the block length the more severe the deterioration.

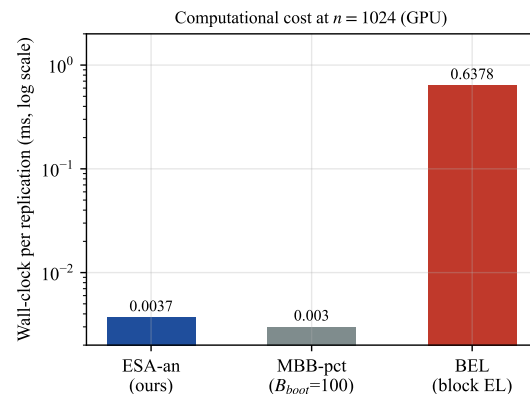


Figure 3. Computational cost ($n = 1024$, GPU, per replication, logarithmic vertical axis). ESA-an is of the same order as the moving block bootstrap and about 170 times faster than the block empirical likelihood.

0.0037 milliseconds, the same order as the moving block bootstrap and roughly 170 times faster than the block empirical likelihood, with no resampling required; see Figure 3.

The scaling law for the optimal block length is shown in Figure 4. In the left panel, $\ell^* / \sqrt{n}$ decreases monotonically in n towards the closed-form constant c , with 0.2727, 0.4875 and 0.7575 for $\lambda = 0.3, 0.6, 0.8$, matching the analytic values at $n = 2^{30}$ to 10^{-4} . The right panel shows the local slope on a logarithmic scale, which rises from 0.42 at $n = 2^{10}$ to 0.4999 at $n = 2^{30}$ but stays within 0.42 to 0.46 throughout the window $n \leq 1024$.

4. Discussion

The results answer the two questions raised in Section 2. Under dependence the tail accuracy of the block self-normalized pivot does not degrade along the scale suggested by the number of blocks: the relative error splits into a bias-and-skewness term of order $n^{-1/2}$ and a floor that decays only with the block length, and the block length at which the two balance is proportional to $\sqrt{n}$. The Gaussian benchmark is therefore not merely slightly inaccurate; its failure grows with the dependence and is worst when the blocks are short.

The floor is the finding we regard as most consequential, since it runs against the intuition that adding blocks increases accuracy. Dependence inflates the long-run variance

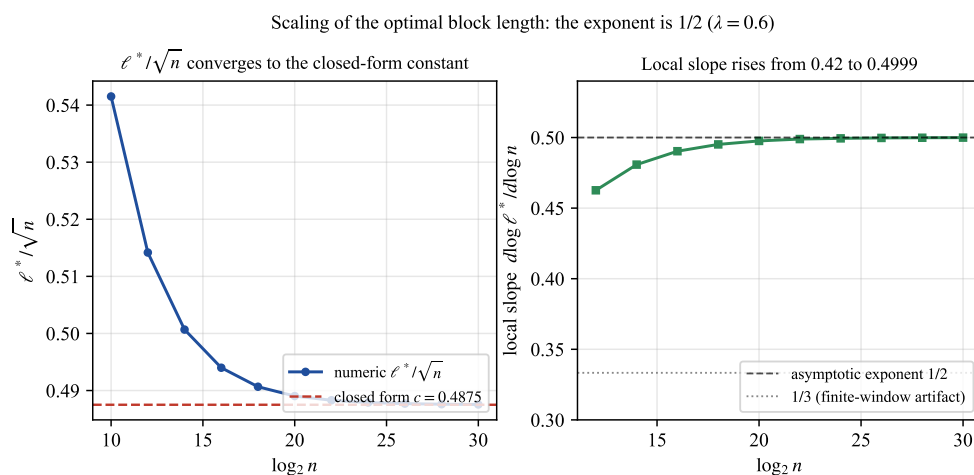


Figure 4. Scaling law for the optimal block length, $\lambda = 0.6$, analytically extrapolated to $n = 2^{30}$. Left: $\ell^* / \sqrt{n}$ converges to the closed-form constant 0.4875. Right: the local slope on a logarithmic scale rises from 0.42 to 0.4999, so the asymptotic exponent is $1/2$ and the $1/3$ seen in a finite window is a spurious exponent.

of the pivot while the benchmark assumes variance one, so no refinement of the number of blocks removes the inflation; the error converges to a limit set by the block length alone. This is why the first-order methods fail exactly where dependence is strong and blocks are short, and why that failure is not a finite-sample artifact that more data repairs. The degradation law is consistent with the three-column error structure that LAHIRI [6] established for sums and means, and it places the quantile case on the $n^{-1/2}$ rather than the $m^{-1/2}$ scale; the ingredient specific to the quantile side is the lattice structure of the pivot, which bounds any continuous approximation of its tail. The scaling exponent should likewise be read with care: a local fit over the accessible sample sizes would suggest an exponent near one third, which the extrapolated analytic slope shows to be spurious.

The results above leave several questions open, and they hold under restrictions that should be stated explicitly. Quantile regression has been listed by [5] as future work, and we treat it only as a natural corollary of the present method rather than a contribution in its own right. The within-block dependence term and the block-length bias term in Theorem 2 cannot be separated in the independent case, and each requires real dependence to be injected before it can be calibrated on its own. When F is discontinuous at q_τ or has a flat density there, the form of the degradation law is not yet determined.

The limits of applicability should also be stated clearly. The optimal block length is confined to the sign region $\lambda > 0$ and $\tau < 1/2$: for $\lambda \leq 0$ the floor becomes a ceiling, and for $\tau \geq 1/2$ the transient and the floor have the same sign, so both cases require separate treatment. The exponential deviation inequality in Assumption (A4) is stronger than polynomial mixing (A1), and for $1 < \beta \leq 2$ the error control of the degradation law still has a gap. The closed-form constant in ℓ^* is the leading-order value obtained by balancing the floor against the transient; the omitted terms of the same order are all pushed down to $O(n^{-1})$ by block averaging and are already absorbed into the remainder of Theorem 2, so they do not alter the scaling law $\ell^* \asymp \sqrt{n}$.

The practical consequence is that the block length ceases to be a free tuning parameter. The closed form evaluates it from the sample size and the fitted dependence, and near that value the coverage error falls from $O(n^{-1/2})$ to $O(n^{-1})$ without a bandwidth beyond the block length itself. This is what the procedure offers the risk measures that motivated the problem, where the block length would otherwise be set by rule of thumb; the same

self-normalization extends to other statistics whose scale can be written as a within-block quadratic form.

5. Conclusions

Under strong mixing, the block self-normalized quantile pivot has an explicit tail-accuracy degradation law and a basis for its block length. The relative tail error is the sum of a bias-and-skewness term of order $n^{-1/2}$ and a variance-inflation floor $E_\infty(\ell) \sim c_\Phi/\ell$: the leading correction is fixed by the sample size, while the only term that survives as the number of blocks grows is the one set by the block length. Balancing the two gives the interior optimum $\ell^* \propto \sqrt{n}$, with a constant that increases with the dependence and diverges as $\tau \rightarrow 1/2$; when $\lambda \leq 0$ the floor becomes a ceiling and no interior optimum exists, which delimits the range of validity. The pivot is lattice-valued, so any continuous approximation of its tail is uniformly off by at least $\Theta(m^{-1}\ell^{-3/2})$, a bound that a piecewise-linear construction attains. At $\ell \asymp \ell^*$ the inverted quantile interval has coverage error $O(n^{-1})$, and in the simulation it holds $95\% \pm 0.15\%$ over sixteen configurations, against 60.1% for the first-order Gaussian baseline at $\lambda = 0.8$, $\ell = 2$. Because the constants enter only through the second and third cumulants of the indicator process, the results are free of F : this carries the two-layer structure that LAHIRI [6] established for sums and means into dependent quantile self-normalization, where the density-free higher-order accuracy proved by SNQESA [5] for independent sampling carries over, and the lattice bound is specific to the quantile side. The form of the degradation law is not yet determined when F is discontinuous at q_τ , and the exponential deviation inequality assumed in (A4) is stronger than polynomial mixing; both are left to subsequent work.

Author Contributions: Conceptualization, methodology, formal analysis, software, and writing: Y.T. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Python scripts that reproduce every table and figure of this study, together with a README describing their use, are provided with the submission and are also available from the corresponding author on reasonable request.

Acknowledgments: During the preparation of this study, the author used DeepSeek-V4.1-Flash for the purposes of translating and language-editing the manuscript and for drafting numerical verification code. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Proofs and Derivations

Appendix A.1. Proof of Density-Freeness

Proof. Since $g_i = \xi_i - \tau$, the pivot T_n is a deterministic function of $\{\xi_i\}$, and the joint law of $\{\xi_i\}$ does not involve F ; hence the distribution of T_n is independent of F .

For the truncation construction, $U_i\tau \sim U(0, \tau)$ and $\tau + U_i(1 - \tau) \sim U(\tau, 1)$. Hence for any x ,

$$\begin{aligned} \mathbb{P}(X_i \leq x) &= \tau \cdot \frac{\min\{F(x), \tau\}}{\tau} + (1 - \tau) \cdot \frac{\max\{F(x) - \tau, 0\}}{1 - \tau} \\ &= \min\{F(x), \tau\} + \max\{F(x) - \tau, 0\} = F(x), \end{aligned}$$

so the marginal distribution is indeed F . Turning to the indicators, take $t = q_\tau$, so that $F(t) = \tau$. If $Z_i = 1$ then $X_i = F^{-1}(U_i\tau)$ and

$$X_i \leq q_\tau \iff U_i\tau \leq F(q_\tau) = \tau \iff U_i \leq 1,$$

which always holds, so $\mathbf{1}\{X_i \leq q_\tau\} = 1 = Z_i$. If $Z_i = 0$ then $X_i = F^{-1}(\tau + U_i(1 - \tau))$ and

$$X_i \leq q_\tau \iff \tau + U_i(1 - \tau) \leq \tau \iff U_i \leq 0,$$

which never holds, so $\mathbf{1}\{X_i \leq q_\tau\} = 0 = Z_i$. For $t \neq q_\tau$ the identity fails, which is the scope stated in Section 2. $\square$

Appendix A.2. Block-Level Cumulants and the Third Moments of the Pivot

Assumption 1(A4) is stronger than the usual mixing conditions; the trade-offs and the sufficient condition for each part are as follows.

Remark A1. (A4) is stronger than (A1) and contains three regularities that must be assumed independently. First, the orders of κ_3 and κ_4 are given by Lemma A4, but determining the transient coefficient requires their sharp remainders: Lemma A3 gives only a remainder $O(m^{-1})$ for κ_3 , which at $\ell \asymp m \asymp \sqrt{n}$ equals the $n^{-1/2}$ order of κ_3 itself and is therefore too coarse to determine the coefficient, so (A4) requires the remainder to drop to $O(n^{-1})$. Second, the variance sharpness $\kappa_2(T_n) = S_{11} + O(m^{-1}\ell^{-1})$, verified against the exact joint law in Appendix A.2 and Table A1. Third, the uniform Edgeworth remainder together with a Bernstein-type deviation inequality. The statement $\kappa_4(T_n) = O(n^{-1})$ is an asymptotic order as $\ell \rightarrow \infty$ with λ fixed; when λ approaches 1 the correlation length $\sim (1 - \lambda)^{-1}$ becomes comparable to ℓ , the decay is numerically slow and the constant is large, but its tail contribution $\frac{\kappa_4/\kappa_2^2}{24}(x^3 - 3x)\phi(x)$ stays at the order 10^{-3} throughout the measurable range, far below $n^{-1/2}$. Bernstein-type inequalities are stronger than polynomial mixing: $\alpha(k) \leq Ck^{-\beta}$ generally yields only polynomial concentration, and exponential concentration requires either $\beta > 2$ together with a Cramér-type condition or geometric (strong) mixing; for $1 < \beta \leq 2$ the error control of the degradation law must be handled separately, and the regular case of this paper assumes geometric mixing. Finally, T_n is lattice-valued, so its Edgeworth expansion carries an oscillatory correction; we use the continuous expansion in Theorem 2 and defer the explicit form of the oscillation to Section 3.3.

Remark A2 (Sufficient Conditions). The three parts of (A4) are guaranteed by a single set of conditions: $\{X_t\}$ is strongly mixing with geometric rate $\alpha(k) \leq C\rho^k$, a moment condition of order $4 + \delta$ holds, and the block variables $V_j = (\tilde{g}_j, \tilde{g}_j^2)$ satisfy the Cramér condition $\sup_{|t| \geq c} |Ee^{itV_1}| < 1$. The block-variable sequence then has mixing coefficients $\alpha_V(k) \leq C\rho^{(k-1)\ell}$, and all the prerequisites of the weakly dependent Edgeworth expansion are met [7,15]. Since $T_n = \sqrt{m}g(\tilde{V})$ is a smooth functional of the block-mean vector ($g(z, w) = z/\sqrt{w}$), and since LAHIRI [6] treats exactly this structure (the smooth function model) under strong mixing and shows that its expansion is a superposition of three series, in powers of $n^{-1/2}$, of $[n/\ell]^{-1/2}$ from the number of blocks, and of the studentizing bias, the sharp κ_3 remainder and the Edgeworth remainder of (A4) follow; the block-scale suppression $S_{12}, \Lambda^{111} = O(\ell^{-1/2})$ then tightens the remainder from $O(m^{-1})$ to $\Theta(m^{-3/2}\ell^{-1/2})$. The variance sharpness of κ_2 also falls within this expansion: under geometric mixing the correlation between nonadjacent blocks decays like ρ^ℓ and the resolvent expansion of $E[T_n^2]$, $\frac{E[A^2]}{m} - \frac{E[A^2C]}{m^2} + \dots$, converges absolutely, with its m^{-1} coefficient given by adjacent-block cumulants and equal to $O(\ell^{-1})$ (the term-by-term estimates in Appendix A.2; numerical verification in Table A1). The Bernstein-type deviation inequality is supplied by MERLEVÈDE, PELIGRAD and RIO [16]; it is also available under polynomial mixing $\alpha(k) \leq Ck^{-\beta}$ with $\beta > 2$, while $\beta \leq 2$ requires an additional truncated moment condition. It should be noted that these sufficient conditions refer to the non-lattice version of the block variables. In the present problem $V_j = (\tilde{g}_j, \tilde{g}_j^2)$ is always lattice-valued, with $\limsup_{|t| \rightarrow \infty} |\hat{\mu}_V(t)| = 1$, so the Cramér condition fails and one must instead use a lattice Edgeworth expansion with oscillatory terms [17]. The lattice branch nevertheless reduces to the non-lattice one: for the smoothed pivot $T_n^s = T_n - \delta V$

with δ equal to one cell width and $V \sim U(0, 1)$ independent, the Cramér condition holds and the expansion applies, while the difference between the exact step function and the smoothed tail is at most $2w_{\max} = O(m^{-1}\ell^{-3/2}) = o(n^{-1})$ (Corollary 1). The $O(n^{-1})$ remainder in the lattice case is thus given jointly by (A4) and the upper bound on the lattice oscillation, and no additional assumption is introduced.

We first record two preliminary facts. Write $\psi_i = Z_i - \tau$, $b_2 = \tau(1 - \tau)$ and $b_3 = b_2(1 - 2\tau)$.

Lemma A1. For a two-state Markov indicator process,

$$\kappa_2(\psi_a, \psi_b) = b_2\lambda^{|a-b|}, \quad \kappa_3(\psi_a, \psi_b, \psi_c) = b_3\lambda^{\max(a,b,c) - \min(a,b,c)}.$$

Proof. Write $p_{01} = \tau(1 - \lambda)$ and $p_{10} = (1 - \tau)(1 - \lambda)$; the transition matrix P has eigenvalues 1 and λ . Decompose $P = \Pi + Q$, where every row of Π equals the stationary distribution $\pi = (\tau, 1 - \tau)$ and $Q = P - \Pi$. From $\Pi P = P\Pi = \Pi$ and $\Pi^2 = \Pi$ we get

$$Q^2 = P^2 - P\Pi - \Pi P + \Pi^2 = P^2 - \Pi = \lambda(P - \Pi) = \lambda Q,$$

so that $P^u = \Pi + \lambda^{u-1}Q$ for $u \geq 1$. Take $e_1 = (0, 1)^\top$; then $Z_i = e_1^\top \xi_i$ and $\text{diag}(\pi)e_1 = \tau e_1$, so that

$$(P^u)_{11} = \text{the first coordinate of } \pi^\top P^u e_1 = \tau + \lambda^{u-1}(P_{11} - \tau) = \tau + \lambda^u(1 - \tau),$$

where we used $P_{11} = 1 - p_{01}$ and $P_{11} - \tau = (1 - \tau)\lambda$. Hence

$$\mathbb{E}[Z_a Z_{a+u}] = \tau (P^u)_{11} = \tau^2 + b_2 \lambda^u, \quad \text{so} \quad \kappa_2(\psi_a, \psi_{a+u}) = \mathbb{E}[\psi_a \psi_{a+u}] = b_2 \lambda^u.$$

For $a < b < c$, multiplying the transition probabilities segment by segment gives

$$\mathbb{E}[Z_a Z_b Z_c] = \tau (P^{b-a})_{11} (P^{c-b})_{11} = \tau^3 + \tau^2(1 - \tau)(\lambda^{b-a} + \lambda^{c-b}) + \tau(1 - \tau)^2 \lambda^{c-a}.$$

Substituting this together with

$$\tau(\mathbb{E}[Z_a Z_b] + \mathbb{E}[Z_a Z_c] + \mathbb{E}[Z_b Z_c]) = 3\tau^3 + \tau b_2(\lambda^{b-a} + \lambda^{c-a} + \lambda^{c-b})$$

into the third joint cumulant $\kappa_3(Z_a, Z_b, Z_c) = \mathbb{E}[Z_a Z_b Z_c] - \tau(\dots) + 2\tau^3$, the constant term gives $\tau^3 - 3\tau^3 + 2\tau^3 = 0$, and the coefficients of λ^{b-a} and λ^{c-b} both reduce to $\tau^2(1 - \tau) - \tau b_2 = 0$ after subtraction, leaving only

$$\tau(1 - \tau)^2 \lambda^{c-a} - \tau b_2 \lambda^{c-a} = \tau(1 - \tau)[(1 - \tau) - \tau] \lambda^{c-a} = b_3 \lambda^{c-a}.$$

Finally, since $\psi_i = Z_i - \tau$ and joint cumulants are invariant under shifts, $\kappa_3(\psi_a, \psi_b, \psi_c) = b_3 \lambda^{\max - \min}$. $\square$

Lemma A2. Let $I_t = \{t\ell + 1, \dots, t\ell + \ell\}$ be the index set of block t and $B_t = \sum_{i \in I_t} \psi_i$. Then

$$\text{cum}(B_{t_1}, B_{t_2}, B_{t_3}) = \sum_{i_1 \in I_{t_1}} \sum_{i_2 \in I_{t_2}} \sum_{i_3 \in I_{t_3}} \kappa_3(i_2 - i_1, i_3 - i_1),$$

and the second-order case follows on replacing κ_3 by κ_2 .

Proof. A joint cumulant is linear in each of its arguments, so expanding B_t over its indices and taking cumulants term by term gives the claim. Neither independence of the B_t nor any condition beyond a possibly nonzero mean of B_t is required. $\square$

Substituting Lemma A1 into Lemma A2 and summing over positions gives

$$S_{11} = \frac{1}{\sigma^2} \sum_{k \in \mathbb{Z}} C_2(0, k), \quad S_{12} = \frac{1}{\sigma^3} \sum_{k \in \mathbb{Z}} C_3(0, k, k), \quad \Lambda^{111} = \frac{1}{\sigma^3} \sum_{a, b \in \mathbb{Z}} C_3(0, a, b),$$

where C_2 and C_3 denote the sums on the right-hand side of Lemma A2, and $\sigma^2 = b_2 R_2(\ell)$ with $R_2(\ell) = \ell + 2 \sum_{d \geq 1} (\ell - d) \lambda^d$. Hence

$$S_{11} = 1 + \frac{2\lambda(1 - \lambda^\ell)}{(1 - \lambda)^2 R_2(\ell)}, \quad S_{12} = \frac{b_3}{\sigma^3} \left[R_3(\ell) + \frac{2M_1(\ell)}{1 - \lambda} \right],$$

$$\Lambda^{111} = \frac{b_3}{\sigma^3} \left[R_3(\ell) + \frac{6M_1(\ell)}{1 - \lambda} + \frac{6\ell\lambda^{\ell+1}}{(1 - \lambda)^2} \right],$$

where $R_3(\ell) = \ell + 6 \sum_{d \geq 1} d(\ell - d) \lambda^d$ and $M_1(\ell) = \sum_{r=1}^{\ell} (2r - 1) \lambda^r$.

Lemma A3. Write $V_j = \bar{g}_j$, $W = \sqrt{m} (V_1, \dots, V_m)^\top$ and $\varepsilon = m^{-1/2}$. From

$$T_n = \sqrt{m} g(\bar{V}_1, \bar{V}_2), \quad g(z, w) = z(1 + w)^{-1/2},$$

expand $T_n = W_1 - \frac{\varepsilon}{2} W_1 W_2 + \dots$ in δ -fashion and compute term by term. Then

$$\kappa_2(T_n) = S_{11} + O(m^{-1} \ell^{-1}), \quad \kappa_3(T_n) = \frac{\Lambda^{111} - 3S_{11}S_{12}}{\sqrt{m}} + O(m^{-1}).$$

Proof. Write $\varepsilon = m^{-1/2}$, $Z_t = B_t/\sigma$ and $\bar{V}_2 = 1 + \varepsilon W_2$. Since $(1 + w)^{-1/2} = 1 - \frac{1}{2}w + \frac{3}{8}w^2 + O(w^3)$,

$$T_n = \sqrt{m} g(\bar{V}_1, \bar{V}_2) = W_1 - \frac{\varepsilon}{2} W_1 W_2 + \frac{3\varepsilon^2}{8} W_1 W_2^2 + O_P(\varepsilon^3).$$

For the second order we have exactly

$$E[T_n^2] = \frac{E[A^2]}{m} - \frac{1}{m} E\left[\frac{A^2 C}{m + C}\right], \quad A = \sum_{t=1}^m Z_t, \quad C = \sum_{t=1}^m (Z_t^2 - 1).$$

The first term is given by Lemma A2 as

$$\frac{E[A^2]}{m} = \frac{1}{m} \sum_{t,s} \Gamma_{t-s}^{11} = S_{11} - \frac{B}{m} + O(m^{-2}), \quad B = \sum_k |k| \Gamma_k^{11}.$$

The inter-block correlation of Z decays like λ^ℓ except between adjacent blocks, and the boundary correlation between adjacent blocks is $\Gamma_1^{11} = \frac{\lambda}{(1-\lambda)^2 R_2(\ell)} = O(\ell^{-1})$, so $B = 2\Gamma_1^{11} + O(\lambda^{2\ell}) = O(\ell^{-1})$. The second term is the source of all $O(m^{-1})$ corrections: expanding it in powers of C/m as $\frac{E[A^2 C]}{m^2} - \frac{E[A^2 C^2]}{m^3} + \dots$, each order satisfies $\frac{E[A^2 C^k]}{m^{k+1}} = \Theta(m^{-1})$, and the $O(1)$ coefficients of adjacent orders are balanced by the identity

$$E[Z^2] = 1 \implies E[Z^2(Z^2 - 1)] = E[(Z^2 - 1)^2]$$

together with the adjacent-block correlations, so that alternating cancellation leaves a kernel of order $O(m^{-1}\ell^{-1})$. Hence

$$\kappa_2(T_n) = S_{11} + O(m^{-1}\ell^{-1}),$$

and for $\ell = O(\sqrt{n})$ this remainder is $O(n^{-1})$, far below the $n^{-1/2}$ order of the floor and the transient. This variance sharpness is recorded as Assumption 1(A4); its order is verified pointwise against the exact joint law: Table A1 reports $\ell m(\kappa_2 - S_{11})$ for $\ell = 2, 4, 8$ and $m \leq 512$ (at $\ell = 16$, $m = 64$ the values are 1.79/3.87/8.12), which rises slowly with m but tends to a finite value, consistent with $O(n^{-1})$.

When the blocks are independent ($\lambda = 0$) the expression can be evaluated completely. From $E[Z_1 Z_2] = E[Z_1 Z_2 C] = 0$ and $E[Z_1 Z_2 C^2] = 2E[Z^3]^2$ (term-by-term computation of the connected sums for $k \leq 3$),

$$E[T_n^2] = 1 + \frac{2E[Z^3]^2}{m} + O(m^{-2}), \quad m_1^2 = \frac{E[Z^3]^2}{4m} + O(m^{-2}),$$

so $\kappa_2(T_n) - 1 = \frac{7E[Z^3]^2}{4m} + O(m^{-2})$. Since $E[Z^3] = O(\ell^{-1/2})$, the coefficient is $O(\ell^{-1})$, consistent with (A4). The exact joint law at $\lambda = 0$ gives $m(\kappa_2 - 1) = 0.31, 0.32, 0.33$ for $\ell = 4$ and $m = 128, 256, 512$, and 0.15 for $\ell = 8$ and $m = 128$; these differ from $\frac{7}{4}E[Z^3]^2$ by at most 10% (for $\ell = 4$ the three cases give 7.9%, 4.0%, 2.0% and for $\ell = 8$, 9.2%), converging as m grows and halving whenever ℓ doubles. In the dependent case the additional adjacent-block cross sums decay like λ^ℓ , or like ρ^ℓ under geometric mixing, and do not change the order.

For the third order, write $T_n = W_1 + \varepsilon A_1 + O_P(\varepsilon^2)$ with $A_1 = -\frac{1}{2}W_1 W_2$. By the multilinearity of κ_3 in each argument,

$$\kappa_3(T_n) = \kappa_3(W_1) + 3\varepsilon \text{cum}(W_1, W_1, A_1) + O(\varepsilon^2).$$

Segment-by-segment summation gives

$$\kappa_3(W_1) = m^{-1/2} \Lambda^{111}, \quad \text{cum}(W_1, W_1, W_1 W_2) = 2 S_{11} S_{12},$$

where $\Lambda^{111} = \sum_{a,b} \text{cum}(Z_0, Z_a, Z_b)$, $\Gamma_k^{12} = \mathbb{E}[Z_0(Z_k^2 - 1)]$ and $S_{12} = \sum_k \Gamma_k^{12}$. The second identity comes from the connected expansion of $W_1 W_2 = (\sum_i V_i)(\sum_j V_j^2)$: pairing one block of W_2 with one W_1 and the remaining two blocks with each other gives $S_{12} \cdot S_{11}$, and the coefficient in the trilinear expansion is 2. Substituting and taking out $\varepsilon = m^{-1/2}$ gives

$$\kappa_3(T_n) = \frac{\Lambda^{111} - 3S_{11}S_{12}}{\sqrt{m}} + O(m^{-1}).$$

For the fourth order and for skewness suppression, Lemma A4 gives $\kappa_3(T_n) = O(n^{-1/2})$ and $\kappa_4(T_n) = O(n^{-1})$. Hence both the fourth-order Edgeworth term $\frac{\kappa_4(T_n)}{24}(x^3 - 3x)\phi(x)$ and the squared-skewness term $\frac{\lambda_3^2}{72}(x^5 - 10x^3 + 15x)\phi(x)$ are $O(n^{-1})$, far below $n^{-1/2}$ at $\ell = O(\sqrt{n})$, and both may be absorbed into the remainder of (A4). $\square$

Lemma A4. Write $Z_t = B_t/\sigma$. Then $S_{12} = O(\ell^{-1/2})$ and $\Lambda^{111} = O(\ell^{-1/2})$, and consequently

$$\kappa_3(T_n) = O(n^{-1/2}), \quad \kappa_4(T_n) = O(n^{-1}).$$

That is, block averaging suppresses the skewness of the pivot by a factor $\sqrt{\ell}$.

Proof. Z_t is a linear combination of ℓ bounded indicators divided by $\sigma = \Theta(\ell^{1/2})$. By Lemma A1 and Lemma A2, the sums defining S_{12} and Λ^{111} are led by $R_3(\ell) = O(\ell)$

Table A1. Numerical verification of the variance sharpness (exact joint law, $\tau = 0.3$). The quantity $\ell m(\kappa_2 - S_{11})$ rises with m but at a decreasing rate, tending to a finite value (the three blocks $\lambda = 0.3/0.6/0.8$ tend to about $2/6/24$), so $\kappa_2(T_n) = S_{11} + O(m^{-1}\ell^{-1}) = S_{11} + O(n^{-1})$, consistent with (A4).

λ	ℓ	$m = 64$	$m = 128$	$m = 256$	$m = 512$
0.3	2	1.98	2.13	2.19	2.22
0.3	4	1.85	2.06	2.16	2.21
0.3	8	1.82	2.11	—	—
0.6	2	4.67	5.33	5.64	5.79
0.6	4	3.83	4.34	4.58	4.70
0.6	8	3.71	4.24	—	—
0.8	2	12.59	17.99	20.43	21.54
0.8	4	10.27	12.52	13.58	14.09
0.8	8	8.43	9.77	—	—

and, upon division by $\sigma^3 = (b_2 R_2(\ell))^{3/2} = \Theta(\ell^{3/2})$, give $S_{12}, \Lambda^{111} = O(\ell^{-1/2})$ (the large- ℓ limits $S_{12}\sqrt{\ell} \rightarrow q_1$ and $\Lambda^{111}\sqrt{\ell} \rightarrow q_1$ are in Appendix A.4). Also $S_{11} = 1 + O(\ell^{-1})$, so $\Lambda^{111} - 3S_{11}S_{12} = O(\ell^{-1/2})$, and substituting into Lemma A3 gives $\kappa_3(T_n) = O((\sqrt{\ell}\sqrt{m})^{-1}) = O(n^{-1/2})$. For the fourth order, $\kappa_4(T_n) = m^{-2}\kappa_4(\sum_t Z_t)$, and the connected quadruple sum grows only like m/ℓ , so $\kappa_4(T_n) = O(m^{-1}\ell^{-1}) = O(n^{-1})$. $\square$

The same verification yields the true remainder of κ_3 . Multiplying $|\kappa_3^{\text{ex}} - \kappa_3^{\text{lead}}|$ by $m^{3/2}\ell^{1/2}$ gives values between 24 and 30, approximately constant, on $(\lambda, \ell, m) = (0.5, 4, 64 \dots 512)$ and $(0.5, 8, 64, 128)$; the remainder is therefore $\Theta(m^{-3/2}\ell^{-1/2})$, which at the balance scale $\ell \asymp m \asymp \sqrt{n}$ is $O(n^{-1})$, better than the $O(m^{-1})$ stated in Lemma A3. This sharpening is exactly what is needed to determine the transient coefficient, and it has been incorporated into (A4).

The conclusion of Lemma A3 can also be written in purely cumulant form. By Lemma A2, the quantities $S_{11}, S_{12}, \Lambda^{111}$ enter only through (κ_2, κ_3) and are independent of F and of the specific structure of the chain. The geometric specialization to a two-state Markov chain is the closed form above; for a general stationary binary process, once (κ_2, κ_3) is known the three constants can be evaluated by the same summation.

Appendix A.3. Derivation of the Degradation Law

Derivation of Theorem 2. By Lemma A3 and the standard Edgeworth expansion,

$$\mathbb{P}(T_n \geq x) = 1 - \Phi(z) + \frac{\lambda_3}{6}(z^2 - 1)\phi(z) + O(n^{-1}), \quad z = \frac{x - m_1}{\sigma_T},$$

where $m_1 = -S_{12}/(2\sqrt{m})$, $\sigma_T = \sqrt{\kappa_2(T_n)} = \sqrt{S_{11}}(1 + O(m^{-1}\ell^{-1}))$ and $\lambda_3 = \kappa_3(T_n)/\kappa_2(T_n)^{3/2}$ is the standardized skewness. The leading term is of order $n^{-1/2}$ and its coefficient carries the factor $b_3 = b_2(1 - 2\tau)$, so the whole term vanishes at $\tau = 1/2$. This is the expression in Theorem 2, in which C_1^{eff} is the leading coefficient relative to the $1 - \Phi$ benchmark and the $O(n^{-1})$ remainder comes from the fourth cumulant and the squared-skewness term omitted in Lemma A3.

The uniformity of the remainder is the content of (A4). Its verification proceeds in two steps: write the difference between the distribution function of T_n and the normal benchmark as an expansion of the characteristic function of W near the origin; then use the Bernstein-type deviation inequality for the block sums to suppress the part of the range in which $\bar{V}_2$ leaves an $m^{-1/2}$ neighbourhood of 1, and expand in Taylor series inside that

neighbourhood, so that the remainder is uniformly $O(n^{-1})$ for $x \in [a, b]$. This exponential deviation inequality is precisely the part of (A4) that is stronger than (A1).

It remains to derive the three branches of $D(\ell)$. The leading term of $D(\ell)$ is determined by the lag-1 autocorrelation of the block sums $Y_j = \sum_{r \in B_j} \psi_r$. Using $\kappa_2(\psi_r, \psi_{r'}) = b_2 \lambda^{|r-r'|}$ from Lemma A1,

$$\text{Var}(Y_1) = b_2 \sum_{r, r'=1}^{\ell} \lambda^{|r-r'|} = b_2 \left[\ell + 2 \sum_{d=1}^{\ell-1} (\ell-d) \lambda^d \right] = b_2 R_2(\ell),$$

$$\text{cov}(Y_1, Y_2) = b_2 \sum_{r=1}^{\ell} \sum_{s=1}^{\ell} \lambda^{\ell+s-r} = b_2 \lambda^{\ell} \left(\sum_{r=1}^{\ell} \lambda^{-r} \right) \left(\sum_{s=1}^{\ell} \lambda^s \right) = b_2 \lambda \left(\frac{1-\lambda^{\ell}}{1-\lambda} \right)^2.$$

Dividing the two and using $R_2(\ell) = \ell(1 + o(1)) \frac{1+\lambda}{1-\lambda}$,

$$\rho_Y(1) = \frac{\text{cov}(Y_1, Y_2)}{\text{Var}(Y_1)} = \frac{\lambda (1-\lambda^{\ell})^2}{(1-\lambda)^2 R_2(\ell)} \sim \frac{\lambda}{\ell(1-\lambda^2)} \sim \frac{C}{\ell},$$

that is, $D(\ell) \sim C/\ell$. When the autocovariance of the block sums decays like $f(k) \sim k^{-\beta}$, the same computation gives $\rho_Y(1) \sim C\ell^{-(\beta-1)}$ for $1 < \beta < 2$ and $\rho_Y(1) \sim C/\ell$ for $\beta > 2$; at $\beta = 2$ the sum $\sum_k k f(k)$ diverges logarithmically, adding a factor $\log \ell$, so $\rho_Y(1) \sim C\ell^{-1} \log \ell$. The three branches combine into the $D(\ell)$ of Theorem 2, and for $\beta \leq 1$ the autocovariance is not absolutely summable and $D(\ell) = O(1)$. These branches give orders only; the constant C is determined by the specific mixing structure, and under geometric mixing it takes the form $\lambda x \phi(x) / [(1-\lambda^2)\Phi(-x)]$ of Lemma 1. The closed-form constant in this paper holds for geometric mixing only. $\square$

Appendix A.4. Derivation of the Optimal Block Length

Proof of Lemma 1. When the block-sum sequence is stationary and mixing, $m \rightarrow \infty$ makes T_n converge in distribution to $N(0, S_{11}(\ell))$, where $S_{11}(\ell)$ is the normalized long-run variance of the block sums. The Gaussian benchmark sets the variance equal to 1, and the discrepancy between the two is $E_{\infty}(\ell)$. When $\lambda > 0$ dependence makes $S_{11}(\ell) > 1$ and $E_{\infty}(\ell) > 0$; when $\lambda < 0$ we have $S_{11}(\ell) < 1$ and $E_{\infty}(\ell) < 0$ (Remark 1). Also, from $R_2(\ell) = \ell(1 + o(1)) \frac{1+\lambda}{1-\lambda}$,

$$S_{11}(\ell) - 1 = \frac{2\lambda(1-\lambda^{\ell})}{(1-\lambda)^2 R_2(\ell)} = \frac{2\lambda}{1-\lambda^2} \cdot \frac{1}{\ell} (1 + o(1)).$$

Expanding $1 - \Phi(x/\sqrt{1+\delta}) = 1 - \Phi(x) + \frac{x}{2} \delta \phi(x) + O(\delta^2)$ for small inflation $\delta = S_{11} - 1$ and substituting into the numerator of (3) gives $E_{\infty}(\ell) \sim c_{\Phi}/\ell$. $\square$

Proof of Lemma 2. Take the large- ℓ limit of the closed forms above. Using $\sum_{d \geq 1} d \lambda^d = \frac{\lambda}{(1-\lambda)^2}$ and $\sum_{d \geq 1} d^2 \lambda^d = \frac{\lambda(1+\lambda)}{(1-\lambda)^3}$,

$$\frac{R_2(\ell)}{\ell} \rightarrow \frac{1+\lambda}{1-\lambda}, \quad \frac{R_3(\ell)}{\ell} \rightarrow 1 + 6 \frac{\lambda}{(1-\lambda)^2} = \frac{1+4\lambda+\lambda^2}{(1-\lambda)^2}, \quad M_1(\ell) \rightarrow \frac{\lambda(1+\lambda)}{(1-\lambda)^2}.$$

Hence $\sigma^3 = (b_2 R_2(\ell))^{3/2} = \Theta(\ell^{3/2})$, while the brackets in S_{12} and Λ^{11} are both led by $\frac{\ell(1+4\lambda+\lambda^2)}{(1-\lambda)^2}$ (the terms M_1 and $\ell \lambda^{\ell+1}$ are $O(1)$ and hence of lower order), so

$$S_{12} \sqrt{\ell} \rightarrow \frac{b_3}{b_2^{3/2}} \left(\frac{1-\lambda}{1+\lambda} \right)^{3/2} \frac{1+4\lambda+\lambda^2}{(1-\lambda)^2} = \frac{1-2\tau}{\sqrt{\tau(1-\tau)}} \cdot \frac{1+4\lambda+\lambda^2}{\sqrt{(1-\lambda)(1+\lambda)^3}} = q_1,$$

and similarly $\Lambda^{111}\sqrt{\ell} \rightarrow q_1$. Hence

$$\text{pred}(\lambda, \ell) = \frac{\Lambda^{111} - 3S_{11}S_{12}}{S_{11}^{3/2}} \rightarrow (q_1 - 3q_1)\ell^{-1/2} = -2q_1\ell^{-1/2},$$

that is, $p_1 = -2q_1$. Substituting into m_1 and λ_3 , and using $\sqrt{m} = \sqrt{n/\ell}$,

$$m_1 = -\frac{S_{12}}{2\sqrt{m}} = -\frac{q_1}{2\sqrt{n}}, \quad \lambda_3 = \frac{\text{pred}}{\sqrt{m}} = \frac{p_1}{\sqrt{n}} = -\frac{2q_1}{\sqrt{n}},$$

neither of which depends on ℓ . The transient is the sum of the bias and skewness terms (tr is the absolute increment of the relative tail deviation, recorded under the same convention in Section 3.2),

$$\text{tr} = \frac{m_1}{\sigma_T}\phi(x) + \frac{\lambda_3}{6}(x^2 - 1)\phi(x) \Big|_{\sigma_T \rightarrow 1} = -\frac{q_1\phi(x)}{2\sqrt{n}} - \frac{2q_1(x^2 - 1)\phi(x)}{6\sqrt{n}} = -\frac{q_1(2x^2 + 1)\phi(x)}{6\sqrt{n}}.$$

□

Proof of Theorem 3. The floor $E_\infty(\ell)$ is a relative quantity while the transient tr is an absolute one, so the two must be placed on a common footing before they can cancel: setting $E_\infty(\ell^*) + \text{tr}(\ell^*, n)/(1 - \Phi(x)) = 0$ and substituting the two expressions,

$$\frac{x\phi(x)\lambda}{(1 - \Phi(x))(1 - \lambda^2)\ell^*} = \frac{q_1(2x^2 + 1)\phi(x)}{6\sqrt{n}(1 - \Phi(x))}.$$

The factor $(1 - \Phi(x))$ appears on both sides and cancels, as does $\phi(x)$, giving

$$\ell^* = \frac{6x\lambda}{(1 - \lambda^2)(2x^2 + 1)q_1}\sqrt{n}.$$

For $\tau = 1/2$ we have $q_1 = 0$ and the equation has no solution; for $\tau > 1/2$ we have $q_1 < 0$ and the closed form returns a negative block length; and for $\lambda \leq 0$ the floor has the opposite sign. All three cases correspond to a transient and a floor that cannot cancel, so no interior optimum exists, and the proof is confined to $\lambda > 0$, $\tau < 1/2$. □

The constant in Theorem 3 is the $\ell \rightarrow \infty$ limit of the transient coefficient. Subtracting the floor $E_\infty(\ell)$ from the relative error of the analytic tail gives the transient, whose coefficient we define as

$$C_1^{\text{eff}}(\ell) = -[\text{err}_{\text{an}}(\ell, m) - E_\infty(\ell)]\sqrt{n}.$$

Table A2 reports the case $\lambda = 0.5$, $x = 2$, $\tau = 0.3$: the quantity $C_1^{\text{eff}}(\ell)$ decreases monotonically in ℓ towards the closed-form value 7.774, with the deviation vanishing at rate $O(\ell^{-1})$, at the same speed as $S_{12}\sqrt{\ell} \rightarrow q_1$ and $\text{pred}\sqrt{\ell} \rightarrow p_1$. The exact joint law at $\ell = 4$, $m = 512$ gives $C_1^{\text{eff}} = 8.459$, differing from the analytic tail by only 0.3%. This comparison shows that the closed-form constant is the limit of the exact pivot and not merely an identity internal to the analytic model.

Appendix A.5. Proof of the Lattice Lower Bound

Lemma A5. Write $S_{ii} = ii - m\ell\tau$ and $Q_{ii,jj} = jj - 2\ell\tau ii + m(\ell\tau)^2$, and set

$$\phi_x(ii) = \frac{S_{ii}^2}{x^2} + 2\ell\tau ii - m(\ell\tau)^2, \quad j_0(ii) = \lfloor 2\ell\tau ii - m(\ell\tau)^2 \rfloor + 1.$$

Table A2. Convergence of the transient coefficient to the closed-form constant ($\lambda = 0.5$, $x = 2$, $\tau = 0.3$, analytic tail). The closed form is $C_1 = q_1(2x^2 + 1)\phi(x)/[6(1 - \Phi(x))] = 7.774$.

ℓ	4	16	64	256	512	2048
$S_{12}\sqrt{\ell}$	2.247	2.201	2.189	2.185	2.184	2.184
$\text{pred}\sqrt{\ell}$	-3.406	-4.139	-4.314	-4.354	-4.361	-4.366
$C_1^{\text{eff}}(\ell)$	8.533	8.110	7.867	7.798	7.786	7.777

Then for $x > 0$,

$$P_{\text{ex}}(x) = \sum_{ii=0}^{m\ell} \sum_{jj=j_0(ii)}^{\lfloor \phi_x(ii) \rfloor} dp[ii, jj].$$

Proof. Since $Q_{ii,jj} = jj - (2\ell\tau ii - m(\ell\tau)^2)$, when $S_{ii} > 0$ we have

$$S_{ii} \geq x\sqrt{Q_{ii,jj}} \iff Q_{ii,jj} \leq \frac{S_{ii}^2}{x^2} \iff jj \leq \phi_x(ii),$$

while for $S_{ii} \leq 0$ the condition can never hold and the contribution is zero. The lower limit $j_0(ii)$ ensures $Q_{ii,jj} > 0$. Summing column by column gives the claim. $\square$

It follows that $\phi_x(ii)$ is continuous and strictly decreasing in x , while the inner sum is a step function of jj ; hence P_{ex} is a nonincreasing right-continuous step function whose jumps occur at $x_{ii,jj} = S_{ii}/\sqrt{Q_{ii,jj}}$, with jump size

$$w(x) = \sum_{\{(ii,jj): x_{ii,jj}=x\}} dp[ii, jj],$$

that is, the sum of dp over all lattice points falling at that x . Different pairs (ii, jj) can give the same x , in which case $w(x)$ is not a single $dp[ii, jj]$.

Lemma A6. Let w be the largest jump size in $[a, b]$ and let x^* be the corresponding jump point. Then for any $h \in C^0([a, b])$,

$$\sup_{x \in [a, b]} |P_{\text{ex}}(x) - h(x)| \geq \frac{1}{2}w.$$

Proof. By right continuity and the existence of the left limit, $P_{\text{ex}}(x^{*-}) - P_{\text{ex}}(x^{*+}) \geq w$. Since h is continuous at x^* ,

$$w \leq |P_{\text{ex}}(x^{*-}) - h(x^*)| + |P_{\text{ex}}(x^{*+}) - h(x^*)| \leq 2 \sup_{[a, b]} |P_{\text{ex}} - h|.$$

$\square$

Proof of Theorem 4. Sharpness is provided by a piecewise-linear construction. Let the jump points in the window be $x_1 < \dots < x_N$ with jump sizes w_i , and take

$$h(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} \left(P_{\text{ex}}(x_i^+) + \frac{1}{2}w_i \right) + \frac{x - x_i}{x_{i+1} - x_i} \left(P_{\text{ex}}(x_{i+1}^+) + \frac{1}{2}w_{i+1} \right), \quad x \in [x_i, x_{i+1}],$$

constant on $[a, x_1]$ and on $[x_N, b]$. This h is continuous. Between jump points P_{ex} is constant while h interpolates linearly, and the deviations at the two ends are $\frac{1}{2}w_i$ and $\frac{1}{2}w_{i+1}$, so

$$|P_{\text{ex}}(x) - h(x)| \leq \frac{1}{2} \max\{w_i, w_{i+1}\},$$

which together with Lemma A6 yields equality.

It remains to determine the order. The dominant jump sizes are given by the probability mass of individual lattice points, and $dp[ii, jj]$ has peak value $1/(2\pi\sigma_S\sigma_Q)$ by the two-dimensional local central limit theorem; coincidences can only make $w_{\max}$ larger and do not change the direction of the lower bound. Let $Y = K - \ell\tau$ with $K \sim \text{Bin}(\ell, \tau)$ and $\sigma^2 = \tau(1 - \tau)$; then

$$\mathbb{E}[Y^2] = \ell\sigma^2, \quad \mathbb{E}[Y^4] = \ell\sigma^2[1 + 3(\ell - 2)\sigma^2],$$

so $\sigma_S^2 = m\ell\sigma^2$ and $\text{Var}(Q) = m\ell\sigma^2[1 + (2\ell - 6)\sigma^2]$. Hence

$$\frac{1}{2\pi\sigma_S\sigma_Q} = \Theta(m^{-1}\ell^{-3/2}),$$

while the boundary curve passes through a region of $O(1)$ relative density inside $[a, b]$, so $w_{\max} = \Theta(m^{-1}\ell^{-3/2})$ and the theorem follows. $\square$

Proof of Corollary 1. Let δ be the width of one lattice cell and take the smoothed pivot $T_n^s = T_n - \delta V$ with $V \sim U(0, 1)$ independent of the data, whose tail function we denote by P_s . Then $P_s(x) = \frac{1}{\delta} \int_{x-\delta}^x P_{\text{ex}}(u) du$, the average of P_{ex} over one cell width. Since P_{ex} is a nonincreasing step function, this average differs from $P_{\text{ex}}(x)$ by at most the total mass of the cells crossed, which is $\leq 2w_{\max}$, so

$$|P_{\text{ex}}(x) - P_s(x)| \leq 2w_{\max}.$$

On the other hand T_n^s is no longer lattice-valued, the Cramér condition holds, and the continuous Edgeworth expansion supplied by (A4) applies to it:

$$|P_s(x) - \Phi_{\text{EE}}(x)| = O(n^{-1}).$$

Adding the two gives $|P_{\text{ex}}(x) - \Phi_{\text{EE}}(x)| \leq 2w_{\max} + O(n^{-1}) = O(n^{-1})$. The lattice term $2w_{\max} \asymp m^{-1}\ell^{-3/2} \asymp n^{-5/4} = o(n^{-1})$ lies below the remainder of the continuous part, so the lattice case has the same order as the continuous case, with only an absorbable oscillatory term in addition.

Numerically ($\lambda = 0.5$, $\tau = 0.3$, $x \in [1.6, 2.4]$, $m \leq 512$), $\sup |P_{\text{ex}} - \Phi_{\text{EE}}|/n^{-1} = 0.38, 0.40, 0.42$ for $\ell = 4$ and 0.41 for $\ell = 8$, a constant times n^{-1} , consistent with a leading term given by the continuous Edgeworth remainder; the lattice term satisfies $w_{\max}/n^{-1} = 0.18, 0.19, 0.22, 0.10$, indeed one order below. The binomial moment formulas used to determine the order in Theorem 4, $\mathbb{E}[Y^2] = \ell\tau(1 - \tau)$ and $\mathbb{E}[Y^4] = \ell\tau(1 - \tau)[1 + 3(\ell - 2)\tau(1 - \tau)]$, were verified term by term from the moment generating function of $K \sim \text{Bin}(\ell, \tau)$. $\square$

References

1. Hall, P.; Sheather, S.J. On the distribution of a studentized quantile. *Journal of the Royal Statistical Society: Series B* **1988**, *50*, 381–391.
2. Hoga, Y.; Schulz, R. Self-Normalized Inference in (Quantile, Expected Shortfall) Regressions for Time Series. arXiv:2502.10065, 2025. <https://doi.org/10.48550/arXiv.2502.10065>.
3. Hounyo, U.; Lin, Y. Two-Way Cluster-Robust Inference for Quantile Regressions. arXiv:2602.16376, 2026.
4. Einmahl, J.H.J.; He, Y. Density-Free Confidence Intervals for Quantiles under Heterogeneity. arXiv:2601.17302, 2026.
5. Hou, J.; Meng, T.; Tian, M. Self-Normalized Quantile Empirical Saddlepoint Approximation. arXiv:2510.24352 [stat.ME], 2025. <https://doi.org/10.48550/arXiv.2510.24352>.
6. Lahiri, S.N. Edgeworth expansions for studentized statistics under weak dependence. *The Annals of Statistics* **2010**, *38*, 388–434. <https://doi.org/10.1214/09-AOS722>.

7. Lahiri, S.N. Asymptotic expansions for sums of block-variables under weak dependence. *The Annals of Statistics* **2007**, *35*, 1324–1350. <https://doi.org/10.1214/009053606000001415>. 686
8. Wang, Y.; Polonik, W. Block Empirical Likelihood for Smooth Functionals. arXiv:1912.07699, 2019. 687
9. Lahiri, S.N.; Sun, S. A Berry–Esseen theorem for sample quantiles under weak dependence. *The Annals of Applied Probability* **2009**, *19*, 1–27. <https://doi.org/10.1214/08-AAP533>. 688
10. Kuffner, T.A.; Lee, S.M.S.; Young, G.A. Optimal Block Bootstrap for Quantiles. arXiv:1710.02537, 2017. 689
11. Athreya, K.B.; Lahiri, S.N. *Measure Theory and Probability Theory*; Springer, 2006. 690
12. Künsch, H.R. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **1989**, *17*, 1217–1241. <https://doi.org/10.1214/aos/1176347265>. 691
13. Liu, R.Y.; Singh, K. Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the Limits of Bootstrap* **1992**, pp. 225–248. 692
14. Owen, A.B. *Empirical Likelihood*; Chapman and Hall/CRC, 2001. 693
15. Götze, F.; Hipp, C. Asymptotic expansions for sums of weakly dependent random vectors. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **1983**, *64*, 211–239. 694
16. Merlevède, F.; Peligrad, M.; Rio, E. Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V: The Luminy Volume*; Institute of Mathematical Statistics, 2009; Vol. 5, *IMS Collections*, pp. 273–292. 695
17. Kolassa, J.; McCullagh, P. Edgeworth series for lattice distributions. *The Annals of Statistics* **1990**, *18*, 981–985. <https://doi.org/10.1214/aos/1176347749>. 696

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content. 704