Automatic stopping with calibrated statistical uncertainties in diffusion Monte Carlo

M. D. Towler
Vallico Academy, Vallico Sotto, 55021 Fabbriche di Vergemoli (LU), Italy[a)]

Automatic stopping at a target statistical error is attractive for unattended and high-throughput diffusion Monte Carlo (DMC) calculations, but stopping when a running error estimate first falls below the target selects downward fluctuations of that estimate. We quantify this effect using ensembles of independent DMC calculations. For a representative all-electron O_2 test, 30.5% of naive stops occur before the correlation-corrected error estimate has matured, in extreme cases underestimating the uncertainty by three orders of magnitude. We introduce a guarded stopping rule that requires the on-the-fly reblocking analysis to pass a validity gate, a safety margin on the target, and persistence of the target condition while the accumulated sample increases by 30%. Matched fixed-length controls show no statistically resolved additional loss of uncertainty calibration due to the stopping decision. In the reference O_2 calculation the median run length is 1.49 ± 0.15 times the ideal fixed length. Tests over target precision, DMC time step, trial-wave-function quality, a second atomic system, and heavy-tailed synthetic data show no new failure mode at the available statistical resolution. The method is implemented in CASINO but is formulated in terms of quantities that can be supplied by other DMC codes.

I. INTRODUCTION

A quantum Monte Carlo (QMC) calculation does not know when it is finished. The quantities it estimates carry statistical error bars that shrink as the inverse square root of the run length, so "finished" is not a state the algorithm reaches but a precision the user must specify.

In normal production work that precision is only one part of a larger calculation design. A QMC practitioner chooses the physical and algorithmic setup, estimates the cost from shorter runs, and allocates effort across such choices as system size, density or geometry grids, twist averaging, DMC time step, walker population, and the statistical length of each calculation. The problem considered here is deliberately the inner one: once a DMC calculation has been specified, how can the code decide for itself that enough statistics have been accumulated to reach a requested standard error on the total energy? This does not address time-step, finite-size, population-control, pseudopotential, localization, or other systematic errors, nor does it optimize the allocation of a global computer budget across multiple calculations.

For unattended and high-throughput workflows, however, this inner stopping problem matters. The obvious automation is to monitor the running estimate of the standard error and stop the first time it falls below the requested target. That rule is statistically dangerous. A stopping decision triggered by a fluctuating uncertainty estimate preferentially selects downward fluctuations of that estimate. In addition, the on-the-fly reblocking analysis can be severely immature early in a run, so an apparently small error bar may simply reflect a failure to have resolved the correlation time.

These are familiar themes in sequential analysis. Fixed-width stopping procedures date to Wald and to Chow and Robbins,[1,2] and asymptotic extensions to correlated simulation output were developed by Glynn and Whitt.[3] Modern Markov-chain Monte Carlo practice uses related batch-means and spectral-variance stopping rules.[4–7] Anytime-valid confidence sequences provide a different solution: they construct a sequence of intervals such that, with a prescribed probability, the true value lies inside all of them simultaneously.[8] The intervals may therefore be inspected repeatedly and the run stopped at a data-dependent time without invalidating their stated coverage. We do not attempt such a construction here. Our aim is instead to retain the on-the-fly reblocked error bar already used in production DMC and ask whether a guarded stopping policy can preserve its fixed-length calibration at finite sample size. The answer is therefore established empirically against fixed-length controls rather than by an anytime-valid coverage theorem.

We first measure the failure of naive stopping directly, using 200 independent all-electron O_2 DMC calculations and fixed-length controls. The failure is not a small correction: 30.5% of runs stop during the immature reblocking regime, some with reported uncertainties wrong by three orders of magnitude. We then construct a guarded stopping rule with three ingredients, using the terminology of the CASINO implementation: (i) a validity gate that prevents any stop before the on-the-fly reblocked error bar is sufficiently well determined; (ii) a margin-adjusted target condition, which increases the probability that the desired precision has genuinely been reached; and (iii) a confirmation window, which requires the target condition to remain satisfied while the accumulated sample grows by a fixed fraction. A drift band provides an additional safeguard against gross pre-existing drift.

We validate the rule against both ordinary fixed-length controls and a more stringent matched-length control ensemble in which each fixed run is stopped at exactly the

[a)] Electronic mail: mdt26@cantab.net

length selected by its paired automatic run. The latter test isolates the effect of the stopping decision from the effect of the resulting distribution of run lengths. The per-run data and analysis scripts are described in the Data Availability statement.

The rule is implemented in CASINO,[9,10] but the statistical construction is not tied to CASINO input syntax, block layout, or reblocking hierarchy. The generic method requires a running estimate of the mean, a correlation-corrected error bar together with an estimate of the uncertainty in that error bar, an estimate of the correlation time or effective sample size, and sufficiently frequent checks. Only the DMC total energy has been validated here; application to other observables or other Monte Carlo methods requires an equivalent on-the-fly uncertainty analysis and separate validation.

II. STATISTICAL UNCERTAINTY IN POPULATION DMC

A. The scalar time series

Let X_t denote the scalar energy estimate accumulated at DMC move t during statistics accumulation. Here one DMC move means one projection time step of the whole configuration (walker) population, and N is the total number of such moves accumulated. In CASINO X_t is the weighted mean local energy over the entire population after the global reduction at that move,

$$X_t = \frac{\sum_i w_{i,t} E_L(\mathbf{R}_{i,t})}{\sum_i w_{i,t}}, \tag{1}$$

with the total population weight attached to that datum, multiplied by the standard population-control reweighting factor when that option is used. In all calculations reported here that reweighting factor is unity (the CASINO default), so each datum is weighted by the total population weight at that move. The running weighted mean of this series is the reported DMC total energy, so the stopping rule monitors the uncertainty of the production estimator itself rather than a separate diagnostic quantity.

Because the scalar series is formed after the global reduction over the whole population, and after all branching up to the preceding DMC move, its serial correlations automatically include persistence of individual walkers, genealogical correlations induced by branching, and correlations introduced by population control. A separate model of those effects is not required by the stopping rule. Population-control *bias* of the mean[11,12] is a different issue: no variance estimator can detect such a bias, and it is outside the scope of the present method (see Sec. VIII).

For a stationary scalar series with point variance s^2, write

$$\mathrm{Var}(\bar{X}_N) \simeq \frac{s^2 N_{\mathrm{corr}}}{N}, \tag{2}$$

where N_{corr} is the CASINO correlation time: the factor by which serial correlation increases the variance of the mean. In the common integrated-autocorrelation convention, $N_{\mathrm{corr}} = 2\tau_{\mathrm{int}}$, and the effective number of independent observations is approximately N/N_{corr}.

B. On-the-fly reblocking

CASINO obtains its correlation-corrected error bar by the reblocking transformation of Flyvbjerg and Petersen,[13] accumulated on the fly as the run proceeds.[14,15] The series is divided into consecutive *reblocking blocks* of B moves, with B a power of two, and the standard error of the mean is estimated from the scatter of the block averages. Successive block lengths are obtained by combining blocks in pairs, and the estimated standard error rises with block length until serial correlations have been absorbed. The production implementation chooses the block length B by the scheme of Lee *et al.*,[15] which rests on Wolff's optimum[16] and selects the smallest available B satisfying

$$B^3 \geq 2N N_{\mathrm{corr}}(B)^2. \tag{3}$$

At block length B, CASINO estimates the correlation time as

$$N_{\mathrm{corr}}(B) = \left[\frac{\hat{\sigma}(B)}{\hat{\sigma}(B=1)}\right]^2, \tag{4}$$

and the quoted N_{corr} is Eq. (4) evaluated at the selected block length. At that level the reblocked error bar is denoted $\hat{\sigma}$. For $n = N/B$ approximately independent block means, CASINO estimates the standard error on the reblocked error bar itself as

$$\mathrm{se}(\hat{\sigma}) = \frac{\hat{\sigma}}{\sqrt{2n-2}}. \tag{5}$$

The relative uncertainty in the reblocked error bar

$$\delta \equiv \frac{\mathrm{se}(\hat{\sigma})}{\hat{\sigma}} \tag{6}$$

therefore depends only on the simple block count $n = N/B$ once B has been selected. The variance estimate itself includes the statistical weights; Eq. (5) nevertheless uses the count N/B.

Equation (5) is the usual standard error of a standard deviation estimated from n independent, normally distributed values. For such values the sample variance $\hat{s}^2$ satisfies $(n-1)\hat{s}^2/s^2 \sim \chi^2_{n-1}$, a distribution of variance $2(n-1)$, so that $\mathrm{Var}(\hat{s}^2) = 2s^4/(n-1)$; propagating this to $\hat{s} = (\hat{s}^2)^{1/2}$ gives, to first order, the relative standard error $1/\sqrt{2(n-1)}$. The factor of two under the square root is thus the variance of a χ^2 variable per degree of freedom, and the same expression is quoted for the uncertainty of the blocked error bar by Flyvbjerg and Petersen [Eq. (28) of Ref. 13].

The particular on-the-fly error-bar estimator is implementation-dependent. A different DMC code could obtain a correlation-corrected error bar by another validated method, for example batch means or spectral analysis. The stopping rule below does not require powers-of-two reblocking or the particular selector of Lee *et al.* It does require some way to determine that the correlation-corrected error bar is trustworthy rather than merely small because the available series is too short to reveal its correlations. An automated test of whether the blocking transformations have converged, such as that of Jonsson,[17] is one form that such a diagnostic could take.

Two different objects are called blocks in what follows. A *reblocking block* is the group of B consecutive moves defined above: it is a construct of the statistical analysis, its length is selected by Eq. (3), and $n = N/B$ counts such blocks. A CASINO block, on the other hand, is a unit of program execution, at whose end the code writes its results and, in the present implementation, tests the stopping condition; its length is set by the user (Sec. V) and is unrelated to B. In the reference calculation of Sec. III, for example, the selected reblocking block length is 512 moves when the validity gate of Sec. IV opens and typically 1024 moves at the end of a run, whereas a CASINO block contains about 3700 moves.

C. Heavy-tailed local-energy statistics

It has long been recognized that the local energies sampled in QMC are not normally distributed and have slowly decaying tails, an observation made explicitly as early as the variational calculations of Pandharipande, Pieper, and Wiringa for helium drops;[18] in electronic-structure QMC the tails follow power laws.[19–21] The resulting non-Gaussianity slows convergence of variance estimates and weakens Gaussian intuition about the relation between sample mean and sample variance. We therefore do not assume that a stopping rule driven by the variance estimate leaves the energy estimate unbiased; this is checked empirically.

Equation (5) should likewise be regarded here as the uncertainty scale used by the on-the-fly reblocking procedure, not as an exact finite-sample confidence interval under an arbitrary heavy-tailed law. Heavy-tail sensitivity is tested separately using synthetic data for which the correlation-corrected variance can be established independently.

III. FAILURE OF NAIVE TARGET-ERROR STOPPING

A. Reference calculation and ensemble design

The reference system is the triplet ground state of all-electron O_2. The trial wave function is a short multi-determinant expansion (12 determinants) in numerical orbitals, multiplied by an optimized Jastrow factor; it is described in Appendix B together with the other trial wave functions used in this work. The DMC time step is 0.003 a.u., the total target weight is 200 (the target population of configurations, summed over all MPI processes, of which there are four), and each calculation is equilibrated for 500 moves (1.5 a.u. of imaginary time) before statistics accumulation begins. The requested standard error is

$$\varepsilon = 5 \times 10^{-3} \text{ a.u.} \tag{7}$$

Two ensembles of 200 independent calculations are used. Arm A uses the naive prototype and evaluates the condition after every DMC move; the guarded implementation introduced later evaluates its condition only at CASINO block ends. In Arm A each run stops at the first such move satisfying

$$\hat{\sigma}(N) \le \varepsilon. \tag{8}$$

Arm B is a fixed-length control of 32,500 DMC statistics-accumulation moves, chosen to match the typical cost of the late branch of Arm A. The two arms use disjoint random seeds. All run lengths below are quoted in the DMC moves defined above. The inverse-variance-weighted mean of the 200 control energies gives the reference energy

$$E_{\text{ref}} = -150.279624 \pm 0.000365 \text{ a.u.} \tag{9}$$

for scoring the independent runs.

Each run i of an ensemble of M independent runs returns a final energy E_i and a reported error bar $\sigma_{\text{rep},i}$, the value of $\hat{\sigma}$ at the end of that run; where no confusion can arise we write simply E and σ_{rep}. Three diagnostics are used. Coverage is the fraction of runs satisfying $|E - E_{\text{ref}}| \le k\sigma_{\text{rep}}$ at $k = 1$ and 2. To compare the reported error bars with the actual run-to-run scatter, we use the calibration ratio

$$R_\sigma = \frac{\langle \sigma_{\text{rep}} \rangle}{\text{SD}(E)}, \tag{10}$$

where $\langle \sigma_{\text{rep}} \rangle = M^{-1} \sum_i \sigma_{\text{rep},i}$ is the mean of the reported error bars and

$$\text{SD}(E) = \left[\frac{1}{M-1} \sum_{i=1}^{M} \left(E_i - \bar{E}\right)^2 \right]^{1/2}, \qquad \bar{E} = \frac{1}{M} \sum_{i=1}^{M} E_i,$$

is the sample standard deviation of the final energies about their own mean. This is the run-to-run scatter: the spread actually observed when the same calculation is repeated with different random seeds, measured without reference to any reported error bar. $R_\sigma = 1$ means that the reported error bars agree on average with that spread; $R_\sigma < 1$ means that they are too small. We also inspect the error in units of its own reported error bar, conventionally called the standardized error or pull,

$$z = \frac{E - E_{\text{ref}}}{\sigma_{\text{rep}}}. \tag{11}$$

Because different runs stop at different lengths, even perfectly calibrated individual error bars do not give exactly $R_\sigma = 1$ when their arithmetic mean is compared with the run-to-run scatter; this small Jensen-inequality effect is quantified below. More importantly, the fixed-length controls themselves need not reproduce textbook Gaussian coverage exactly. The relevant test is whether automatic stopping introduces an *additional* calibration defect beyond the uncertainty estimator used in ordinary fixed-length DMC.

Arm B has one- and two-standard-error coverages of 64.0±3.4% and 94.5±1.6%, respectively, and $R_\sigma = 0.934$.

B. Two distinct failure modes

Figure 1 shows that the naive stopping distribution is strongly bimodal. The two modes correspond to two distinct consequences of naive stopping:

(i) *Immature-correlation stops.* The run stops within its first few hundred moves, before the reblocking analysis has resolved the correlation time, and reports an error bar that can be too small by orders of magnitude.

(ii) *First-crossing selection.* The run stops after the reblocking analysis has matured, but on a downward fluctuation of the error estimate, so that the reported error bar is too small by a modest but systematic amount.

We quantify each in turn.

(i) Immature-correlation stops. Sixty-one of 200 runs (30.5%) stop after only 4 to 361 moves, before the on-the-fly reblocking analysis has identified the correlation time. These are not ordinary downward fluctuations of a mature variance estimator. The estimator is systematically blind to correlations that have not yet had time to appear. The most extreme run reports $\sigma_{\rm rep} = 2.3\times10^{-6}$ a.u., roughly three orders of magnitude smaller than the requested target. The one- and two-standard-error coverages of this early group are both $6.6 \pm 3.2\%$.

(ii) First-crossing selection. The remaining 139 runs stop much later. Of these, 137 exceed 10^4 moves, and the median stopping point of the late group is 31,447 moves. Their uncertainties are no longer catastrophically wrong, but first-crossing selection remains visible: $R_\sigma = 0.814$ compared with 0.934 for the fixed-length controls, and one-standard-error coverage is $55.4 \pm 4.2\%$.

Across all 200 naive runs the one- and two-standard-error coverages are $40.5 \pm 3.5\%$ and $66.0 \pm 3.3\%$, and $R_\sigma = 0.092$. The naive rule is therefore unusable. The energy itself shows no statistically resolved bias within the late-stop branch: its ensemble mean agrees with $E_{\rm ref}$ within 0.4 mHa. At the resolution of these ensembles, the dominant damage is to the reported uncertainty rather than to the central energy estimate.

Relation to earlier stopping rules. Equation (8) is the fixed-width, or absolute-precision, sequential rule of the literature cited in Sec. I, which stops the first time the estimated half-width of a confidence interval falls below the target,[2,3] applied to the error bar that production DMC already reports; Glynn and Whitt note that the unmodified rule can stop far too early when the variance estimator is badly behaved at small times, and the termination of a QMC calculation once a specified convergence is attained was one of the stated purposes of the on-the-fly error analysis of Kent *et al.*[14] Arm A measures how that rule performs here when nothing guards it. Fixed-width practice in Markov-chain Monte Carlo guards it by requiring a minimum simulation effort, chosen in advance, before the rule may fire,[4,6] and effective-sample-size rules impose a floor of a related kind.[5] Such rules have not been evaluated here: the late-stopping group (ii) is a subset selected by having avoided an early crossing, not an ensemble run under a minimum-length rule, and its undercalibration shows only that first-crossing selection persists once the analysis has matured. The rule of Sec. IV sets its minimum length from the data, in units of the measured correlation time, and adds the requirement of persistence over sample growth; whether that requirement leaves any calibration defect is not argued but tested, against fixed-length controls, in Sec. VI.

IV. A GUARDED STOPPING RULE

The two failures above require different remedies. The immature-correlation failure must be prevented outright; the first-crossing selection effect must be allowed to decay by accumulating substantial new data after the apparent crossing.

A. Validity gate

No target test is permitted until the validity gate has opened. In the present implementation this requires four conditions on the on-the-fly reblocking analysis. First, the reblocking analysis must have found a plateau. Second, the standard error on the reblocked error bar must satisfy

$$\delta = \frac{\mathrm{se}(\hat\sigma)}{\hat\sigma} \le 0.15. \tag{12}$$

With Eq. (5) this is equivalent to at least

$$n \ge 1 + \frac{1}{2(0.15)^2} \simeq 23.2 \tag{13}$$

approximately independent reblocking blocks. Third, the selected block length must have at least one longer computable block length above it, so that the plateau was chosen rather than imposed by running out of block lengths. Fourth, the run must span at least

$$N \ge 100 N_{\rm corr}, \tag{14}$$

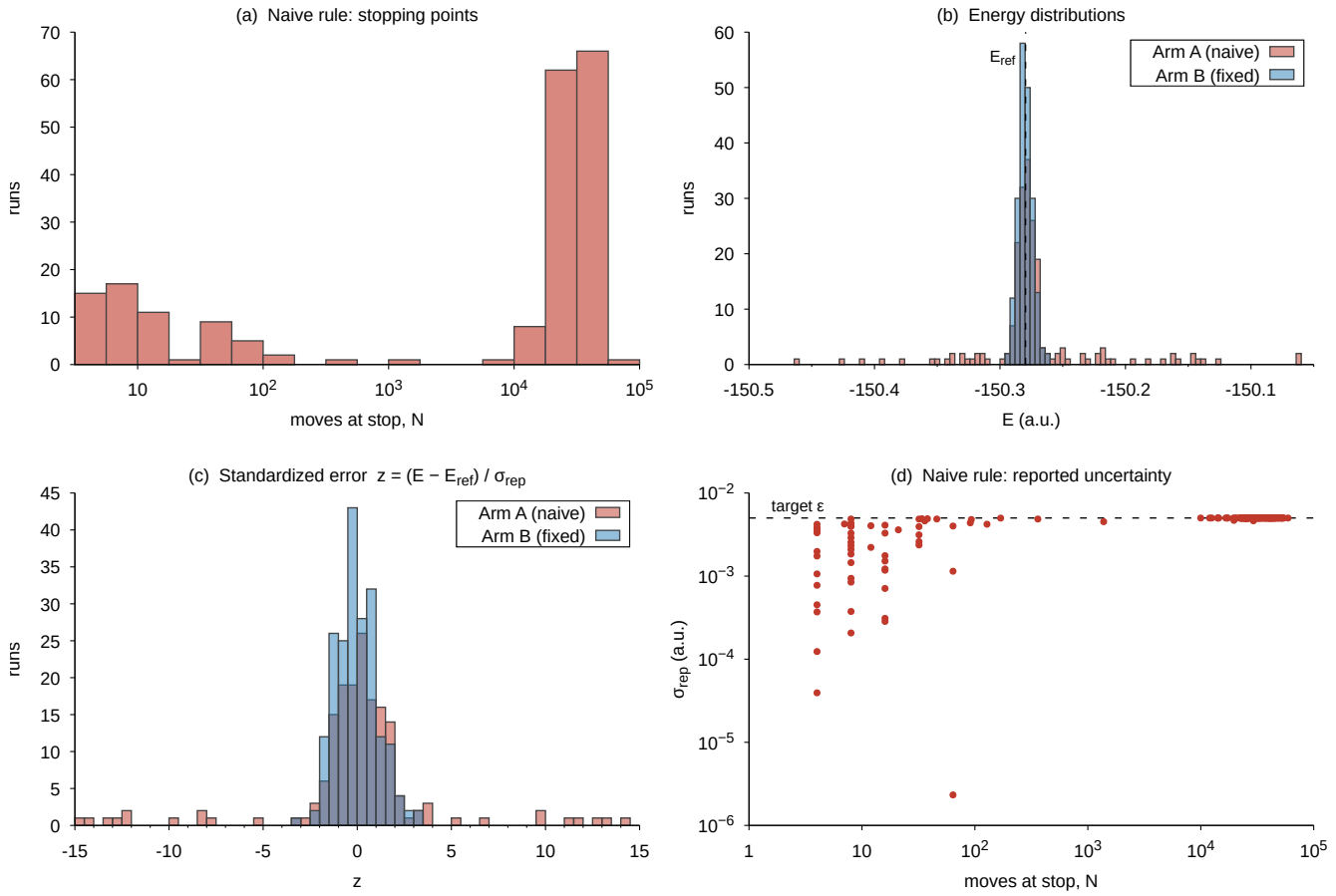


FIG. 1. Naive first-crossing stopping (Arm A, 200 runs) compared with fixed-length controls (Arm B, 200 runs) for all-electron O_2. (a) Distribution of the stopping points of Arm A, in bins of width 0.25 in $\log_{10} N$, showing the early immature-correlation mode. (b) Energy distributions; the dashed line marks E_{ref}. (c) Standardized errors $z = (E - E_{\mathrm{ref}})/\sigma_{\mathrm{rep}}$; 33 early stops of Arm A have $|z| > 15$ and lie outside the range shown. (d) Reported uncertainty of Arm A against stopping point; the dashed line is the requested target ε.

or roughly one hundred effective independent observations in the convention used here. For the present block selector the last condition is normally redundant with Eq. (12) and serves mainly as an additional safeguard against anomalous block selection. A criterion of the same kind was proposed for the analysis of completed runs by Lee *et al.*, who regard a reblocking estimate as reliable only when the selected block length is below $N/50$, that is, when at least 50 reblocking blocks are available;[15] the gate here asks for about 23, but it is only the first of three conditions. The thresholds are dimensionless and tied to the measured correlation structure rather than to a fixed number of DMC moves.

A replay of 300 production histories emphasizes what the validity gate does and does not establish. At the instant the gate first opens, $\hat{\sigma}$ is still only about 0.75–0.79 of the reference error-bar scale on average, and 40–46% of openings are more than 30% low. The gate removes the grossly immature regime; it does not by itself certify the final calibration.

B. Margin-adjusted target condition

Once the validity gate is open, the instantaneous condition is

$$\hat{\sigma} + \kappa\, \mathrm{se}(\hat{\sigma}) \le \varepsilon, \qquad \kappa = 1. \tag{15}$$

The role of this margin is limited. In the regime observed here, changing N mainly rescales both the reported and actual errors as $N^{-1/2}$ without appreciably changing their standardized distribution. Lowering the threshold therefore moves first crossing to a later point but does not remove the first-crossing calibration bias.

What it does improve is the probability that the reference correlation-corrected error bar at the stopping length is genuinely at or below ε. We refer to this probability as target attainment. Because the stopping time is data-dependent, no fixed-sample Gaussian confidence interpretation is attached to κ; target attainment is measured directly in Sec. VI.

C. Persistence over sample growth

The first-crossing selection effect is tied to a downward fluctuation present when the condition first becomes true. Consecutive checks of a cumulative estimator do not provide independent confirmation if little new data have been added. The natural scale is therefore sample growth rather than simply the number of successive checks.

Let N_0 be the number of moves at the first check, after the validity gate has opened, that satisfies Eq. (15). The condition must remain satisfied until

$$N \geq (1+g)N_0, \qquad g = 0.30. \tag{16}$$

The original fluctuation is then diluted by substantial fresh data. A simple scaling argument suggests that the reported-to-actual uncertainty ratio should improve by roughly $\sqrt{1+g}$. Applied to the late branch of the naive rule, $0.814\sqrt{1.3} \simeq 0.93$, close to the measured guarded value. This is a mechanism estimate rather than a theorem; calibration is established by the ensembles below.

D. Drift band

During the confirmation window the running mean is required to remain within $\pm 2\hat{\sigma}$ of the mean at the beginning of the window, with $\hat{\sigma}$ evaluated at the current check. Exceeding this tolerance resets the confirmation. Synthetic tests show that this is useful against gross drift already present when the window begins but has poor power against changes that begin inside the window. The method therefore assumes that the production accumulation is stationary; the drift check is not an equilibration detector.

E. Generic algorithm

Figure 2 summarizes the stopping logic. At each check the calculation updates the running mean and its correlation-corrected error bar. If the validity gate is closed, the target is not tested. Once the gate is open, the margin-adjusted condition must first become true and then remain true while the sample grows by 30%; until that growth is complete the run continues with the window kept. A failure of the target condition or the drift-band check resets the window.

The statistically important confirmation is the fractional growth in Eq. (16). CASINO additionally requires the condition to hold at three consecutive checks. Because CASINO makes its checks at the ends of CASINO blocks (Sec. V), this is a requirement of three consecutive CASINO block ends. It does not refer to the reblocking blocks of Sec. II, which enter the rule only through $\hat{\sigma}$, δ, and N_{corr}. The requirement is a conservative engineering safeguard; the statistical protection against first-crossing selection comes from requiring the extra 30% of data. It is inactive when checks are frequent compared with the growth window, but can add conservative overhead when CASINO blocks are long.

V. CASINO IMPLEMENTATION

CASINO supplies the statistical quantities above from its on-the-fly powers-of-two reblocking analysis. At each DMC move the population-weighted local-energy mean is added to the analysis after reduction over all MPI processes. The stopping criterion is checked at the ends of CASINO blocks.

CASINO block ends are not a statistical requirement; they are simply where CASINO performs the check. Another code may perform the same check at any sufficiently frequent point at which the running mean and a correlation-corrected error bar are available.

The fixed internal parameters are the safety-margin multiplier $\kappa = 1$, the fractional sample-growth requirement $g = 0.30$, the maximum relative uncertainty of the reblocked error bar $\delta_{\max} = 0.15$, the requirement that the run span at least $100N_{\mathrm{corr}}$, a drift tolerance of $\pm 2\hat{\sigma}$, and three consecutive passing checks (CASINO block ends) during confirmation. These are fixed in the code rather than supplied as input keywords.

The user selects the stopping rule with `STOP METHOD`. For target-error stopping, `TARGET ERROR` gives the requested error bar. `BLOCK TIME` sets the length of a CASINO block and hence the interval between checks; it is specified in minutes of master-process CPU time, not wall time. `STOP TIME` gives the master-process CPU-time ceiling for target-error stopping and the available CPU-time budget for the budget-limited mode of Appendix A. Long CASINO blocks make the stopping point coarser and can extend a run substantially beyond the 30% growth requirement.

Restarting preserves the accumulated statistics, including N, $\hat{\sigma}$, the selected reblocking length, and the correlation estimate, but not progress through the current confirmation window. A restarted calculation therefore starts a fresh confirmation window from the restored statistics. This normally delays confirmation; the exact CASINO block ends and drift reference also change on restart. These details are properties of the present CASINO implementation rather than requirements of the statistical rule.

The target-error implementation also contains a CPU-time safeguard. Once the validity gate has opened, CASINO estimates the total CPU time needed to reach the target and compares it with `STOP TIME`. To avoid abandoning a run because of one noisy estimate, it uses the median of the last five projections, one made at each CASINO block end, and requires three consecutive over-budget results before stopping. This safeguard is separate from the statistical argument for target-error stopping.

The present CASINO implementation is available for

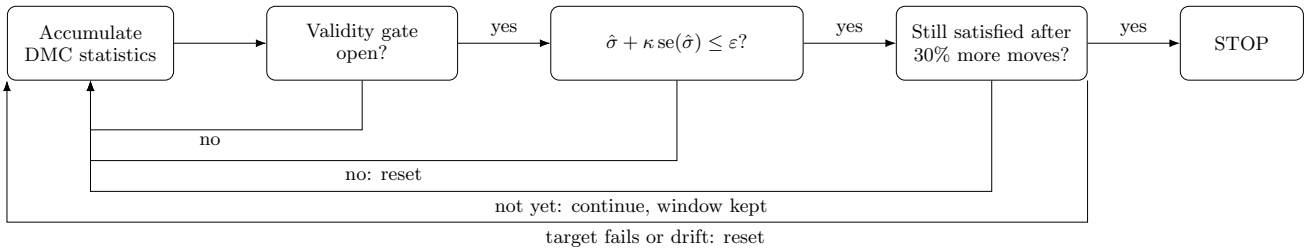


FIG. 2. Code-independent logic of the target-error rule. The validity gate prevents stopping on an unresolved reblocking analysis; the margin makes it more likely that the requested error bar has genuinely been reached; persistence over a fixed fractional increase in sample size suppresses first-crossing calibration bias. The check may be made whenever the required statistical quantities are available, provided it is made sufficiently often.

DMC but not for pure VMC. The reason is practical rather than statistical, and specific to the present code. First, the control of a VMC run whose length is not fixed in advance is unfinished in the CASINO VMC driver: the loop bounds for such a run exist, but the branch that would end it on the error bar has not been written, and the input is refused for any stopping method other than a fixed number of moves rather than allowed to run indefinitely. Second, the VMC on-the-fly reblocking analysis does not receive the series on which the rule was validated: its entries are unweighted averages over the MPI processes, so the variance of an entry depends on the number of processes, whereas the DMC datum of Eq. (1) is the weighted average over a population whose total size is fixed however it is distributed; the thresholds of the validity gate would have to be validated afresh for the VMC series. We know of no statistical obstacle to applying the same rule to a VMC energy series, but it has been neither implemented nor validated, and it is outside the present paper.

VI. VALIDATION

A. Paired validation ensemble

A third ensemble of 200 O_2 calculations (Arm C) uses the same random seeds as the naive ensemble A but applies the guarded rule with the values above. The underlying stochastic trajectory is therefore reproduced identically up to Arm A's stopping point, although Arm C tests its criterion only at CASINO block ends and continues when required.

Figure 3 compares the guarded ensemble with Arms A and B. No guarded run stops in the immature regime [Fig. 3(a)]. The earliest stopping point is 25,332 moves, well beyond the approximately 12,000-move validity-gate threshold of the reference calculation (the $\delta = 0.15$ threshold corresponds to about 23.2 reblocking blocks, with $B \simeq 512$ here). The 61 trajectories that the naive rule abandoned within its first few hundred moves now stop at a median of 62,691 moves, indistinguishable from the median of 62,537 moves for the other 139 [Fig. 3(b)].

The principal calibration result is

$$R_\sigma = 0.935 \tag{17}$$

for the guarded ensemble compared with 0.934 for the ordinary fixed-length controls. Bootstrap resampling of the 200 runs gives an approximate 95% half-width of 0.05 for the guarded ratio, so the agreement in the third decimal place is not significant.

Coverage is likewise consistent: the guarded ensemble gives $67.5 \pm 3.3\%$ at one standard error and $93.0 \pm 1.8\%$ at two, compared with $64.0 \pm 3.4\%$ and $94.5 \pm 1.6\%$ for Arm B; the two distributions of standardized errors are compared in Fig. 3(c). The result to emphasize is not that either ensemble exactly reproduces textbook Gaussian coverage, but that the stopping decision introduces no statistically resolved additional miscalibration.

B. Matched-length fixed-run control

The strongest test isolates the stopping decision itself. A fresh ensemble of 200 fixed-length runs is generated, with run i stopped after exactly the number N_i of moves selected by guarded run i. No stopping criterion is consulted. Table I compares the two ensembles.

The matched-control minus guarded coverage differences are -4.5 ± 4.7 percentage points at one standard error and $+1.0 \pm 2.5$ points at two. No calibration effect attributable to the stopping decision is statistically resolved. The residual $R_\sigma < 1$ is therefore primarily a property of the underlying reblocking estimator rather than of automatic stopping. The variation in stopping lengths accounts for only about 1.5 percentage points of this deficit. At typical guarded stopping points δ is about 0.09 (it is 0.15 by construction at the resolution threshold, and about 0.13 at the fixed length of Arm B). The automatic block selector balances the systematic error of the reblocked estimate against this statistical one,[15,16] so a residual downward bias of the reblocked error bar at the selected block length is expected, of a size comparable with the deficit observed; the matched-length control, rather than this expectation, carries the calibration conclusion.

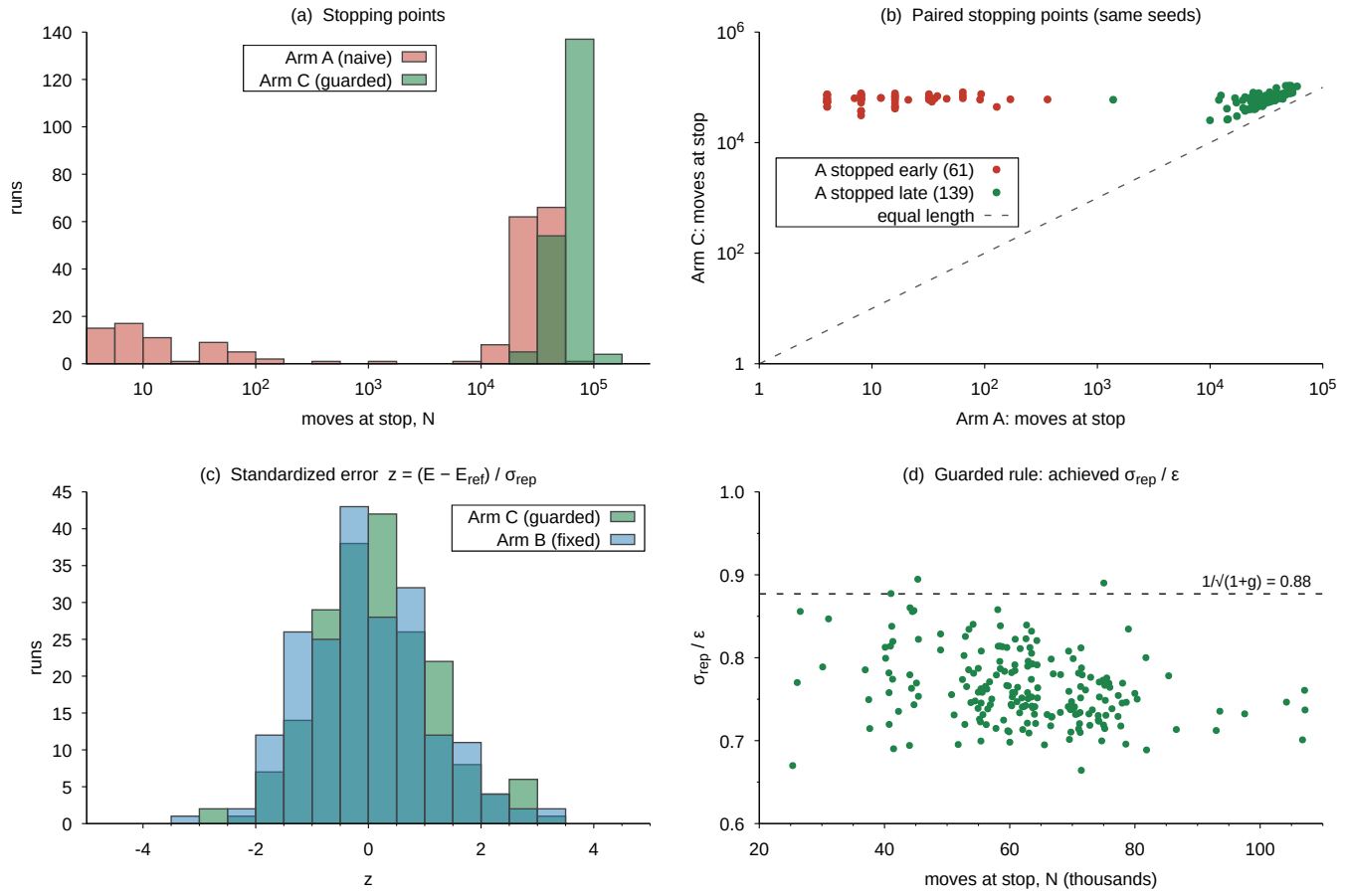


FIG. 3. Guarded stopping (Arm C) compared with the naive rule and the fixed-length control. (a) Stopping-point distributions, in bins of width 0.25 in $\log_{10} N$; the immature mode of Arm A is absent from Arm C. (b) Paired stopping points of the runs with the same random seeds, showing that the trajectories abandoned early by the naive rule continue under the guarded rule to the same lengths as the others; the dashed line marks equal length. (c) Standardized-error distributions for Arm C and fixed-length Arm B. (d) Achieved $\sigma_{\mathrm{rep}}/\varepsilon$ against stopping point for Arm C. The dashed line marks $1/\sqrt{1+g} \simeq 0.88$, the ratio that would result if the error bar equaled the target when the confirmation window opened and then decreased smoothly as $N^{-1/2}$ over the minimum sample growth of 30%; most observed ratios are lower, because the window opens only once the margin-adjusted condition of Eq. (15) is met and the stop then falls at the end of a whole CASINO block, and the fluctuations of the error bar put a few above the line (Sec. VI).

C. Target attainment

To ask whether the requested precision has genuinely been reached, define a reference error-bar scale

$$\sigma_{\mathrm{ref}}(N) = \frac{A_\sigma}{\sqrt{N}}, \tag{18}$$

with A_σ estimated from the fixed-length control ensemble. In 90% of guarded stops, $\sigma_{\mathrm{ref}}(N_{\mathrm{stop}}) \le \varepsilon$. The binomial 95% confidence interval is 85–94%, conditional on the control-calibrated A_σ. Rescoring the archived stopping lengths after shifting A_σ by two standard deviations gives the sensitivity range 83–97%; no additional fit is involved.

Synthetic tests isolate the effect of the margin. At a comfortable target both $\kappa = 0$ and $\kappa = 1$ already saturate target attainment (98.8% and 99.8%). At a marginal target, changing κ from 0 to 1 raises target attainment from 82% to 92% among runs that stop within the synthetic horizon; the corresponding firing fractions are 86% and 57%, so the improved conditional attainment is purchased partly by refusing or delaying marginal cases. Coverage remains statistically unchanged. This confirms the intended division of labor: the margin improves target attainment, whereas the sample-growth window controls stopping-induced calibration bias.

D. Computational overhead

Three cost comparisons should not be conflated. The guarded runs have a paired median length 1.80 times

TABLE I. Guarded Arm C and the matched-length fixed-run control. Brackets give bootstrap 95% intervals for the root-mean-square (RMS) standardized error. The coverages of the guarded ensemble are scored against the weighted reference energy of Sec. III; those of the matched control, and both RMS values, against the unweighted mean of the control energies, -150.279706 ± 0.000403 a.u., which differs from it by 0.00008 a.u. (0.2 standard errors). On the weighted reference the matched-control coverages are $62.0 \pm 3.4\%$ and $93.5 \pm 1.7\%$ and the RMS z of Arm C is 1.075.

Quantity	Guarded C	Matched fixed
1σ coverage (%)	67.5 ± 3.3	63.0 ± 3.4
2σ coverage (%)	93.0 ± 1.8	94.0 ± 1.7
R_σ	0.935	0.934
RMS z	1.078 [0.97,1.19]	1.080 [0.98,1.17]

that of the late branch of the naive rule, but those late naive stops are themselves selected and undercalibrated, and the ratio also includes the change from checks after every move in Arm A to checks only at CASINO block ends in Arm C. The mean achieved uncertainty is $\langle \sigma_{\rm rep}/\varepsilon \rangle = 0.762$ [Fig. 3(d)]. We define the delivered-precision factor of a run as

$$P_{\rm rep} = \left(\frac{\sigma_{\rm rep}}{\varepsilon} \right)^{-2} \tag{19}$$

and summarize it by its median over an ensemble, the convention of Table II; for Arm C the median is $P_{\rm rep} = 1.75$ (the inverse square of the mean ratio is 1.72). Two idealized reference values help in reading such numbers. If the margin-adjusted condition of Eq. (15) became an equality at a check and the run stopped there, its delivered-precision factor would be $(1 + \kappa\delta)^2$; since $\delta \le \delta_{\rm max}$ whenever the gate is open, this boundary value cannot exceed $(1 + \kappa\delta_{\rm max})^2 \simeq 1.32$, and it is 1.27 for $\delta \simeq 0.13$, typical of a check made soon after the gate opens, and about 1.2 for the $\delta \simeq 0.09$ found at the stopping points of Arm C. A real run overshoots the boundary, because the condition is tested only at checks and may already hold when the gate opens, and then runs on through the confirmation window. Similarly, if the error bar equaled the target when the window opened and then decreased smoothly as $N^{-1/2}$ over the minimum sample growth, the achieved ratio would be $\sigma_{\rm rep}/\varepsilon = 1/\sqrt{1+g} \simeq 0.88$; most observed ratios lie below this value, because the window opens only once the margin-adjusted condition holds and the stop is delayed to the end of a whole CASINO block, while the fluctuations of the error bar put a few above it. The most transparent reference is an ideal fixed-length run whose required length is known in advance from Eq. (18):

$$N_{\rm ideal} = \left(\frac{A_\sigma}{\varepsilon} \right)^2 \simeq 4.21 \times 10^4. \tag{20}$$

With median guarded length $N_{\rm stop} = 62568$,

$$\frac{N_{\rm stop}}{N_{\rm ideal}} = 1.49. \tag{21}$$

The uncertainty in A_σ alone gives approximately ± 0.15 in this ratio. We therefore describe the reference overhead as roughly 1.5 times the ideal fixed length, not as a universal constant. The dominant *implementation-specific* contribution is quantization of the 30% confirmation window to whole CASINO blocks, which always moves the stop later. In run-length accounting relative to the threshold, the margin and the 30% growth window are the larger designed contributions; quantization to CASINO blocks is an additional implementation-specific factor.

E. Synthetic known-truth tests

Synthetic correlated series provide tests for which the correlation-corrected scatter can be established independently. We use Gaussian first-order autoregressive (AR(1)) series with nominal CASINO correlation time $N_{\rm corr} = 30$. A heavy-tailed version is made from the same underlying Gaussian series by transforming each value to have a unit-variance Student-t_3 distribution. This nonlinear transformation preserves the rank dependence but not the ordinary linear autocorrelation function, so the transformed series' correlation time is measured directly. The resulting factor is 27.3 ± 1, with block-size estimates 27.5 and 27.1. The corrected heavy-tail result changes by less than half a percentage point across this range.

For mature guarded stopping, replacing Gaussian marginals with t_3 marginals lowers the calibration ratio by about 2.6 percentage points while moving coverage by only about 2.5 points. Fixed-length analysis inherits essentially the same effect. The same synthetic machinery reproduces the catastrophic naive-stopping failure: with frequent checks, 72% of naive stops underestimate the reference uncertainty by at least a factor of two, some by order 10^3, and coverage falls to 24%. Adding the validity gate removes this class in 400 test runs.

A 27-cell sensitivity map over $g \in \{0.2, 0.3, 0.5\}$, $\delta_{\rm max} \in \{0.10, 0.15, 0.20\}$, and drift-band width $\in \{1.5, 2, 3\}\hat{\sigma}$ finds no immature-stop failures in any cell. Calibration varies monotonically with the sample-growth fraction as expected; no statistically resolved dependence on drift-band width is seen in the stationary tests.

VII. ROBUSTNESS ACROSS TARGETS, TIME STEPS, AND SYSTEMS

The guarded rule is next tested without retuning in eleven further ensembles, which we call cells. Nine form a 3×3 grid for the O_2 system of Sec. III: requested standard errors ε of 2, 5, and 10×10^{-3} a.u. crossed with DMC time steps δt of 0.0015, 0.003, and 0.006 a.u., that is, 0.5, 1, and 2 times the reference value; a grid cell is labeled

by these two numbers, so that "5 / 1×" is the reference calculation repeated. Cell E uses the reference target and time step with a deliberately degraded O_2 trial wave function, in which the Jastrow factor is reduced to its electron-electron term (see below and Appendix B); cell F is an all-electron Ne atom with $\delta t = 0.002$ a.u. and $\varepsilon = 4.3 \times 10^{-3}$ a.u., the same relative precision as the reference target. Each cell has 30 independent guarded runs, the reference cell 100; all use a total target weight of 200, and the DMC equilibration is 1.5 a.u. of imaginary time for O_2 and 4 a.u. for Ne. Because the reference energy and the calibration of the underlying reblocking estimator change with time step and system, each cell is scored against a fixed-length control ensemble of 100 runs: there is one control ensemble per time step of the grid, shared by the three targets at that time step, and one each for E and F, each run for the median stopping length of the guarded cell it serves (for the grid, of the $\varepsilon = 5 \times 10^{-3}$ a.u. cell at that time step) and supplying the reference energy against which that cell is scored. The campaign thus contains 400 guarded runs and 500 controls; Fig. 4 and Table II give the results. No guarded run stops below its own validity-gate threshold, no run is declared unreachable, and no step cap is encountered.

At one standard error, 9 of the 11 guarded cells agree with their own fixed-length controls, the difference in coverage being less than twice its binomial uncertainty; all 11 agree at two standard errors. The two one-standard-error deviations occur in opposite directions and are both $n = 30$ cells. The fixed-length controls themselves show why such cell-specific comparison is necessary: at half the reference time step, for example, the control has one-standard-error coverage $56.0 \pm 5.0\%$.

The degraded O_2 wave function of cell E is obtained from the reference one by removing every term of the Jastrow factor except the isotropic electron-electron term u, whose parameters are then reoptimized by variance minimization; the determinantal part is unchanged. The Jastrow factor is thus truncated rather than omitted, and no cusp condition is violated: the electron-electron cusps are imposed by the u term, and the electron-nucleus cusps are a property of the numerical orbitals. Removing the explicit electron-nucleus and electron-electron-nucleus correlation raises the variational energy by about 0.15 a.u. and increases the local-energy variance from 2.62 to 5.66 a.u.^{2} (Appendix B). Its standardized tails are in fact lighter than those of the optimized wave function, so this is a high-variance rather than a heavy-tail stress test. No failure of the data-scaled validity gate is observed.

The Ne calculation exposes an implementation boundary rather than a statistical failure. Ne is so cheap per DMC move that a single CASINO block contains a large fraction of the required sample: more than 20,000 moves, against about 3700 for O_2, while the correlation time is only about 8 moves. The requirement for three consecutive CASINO block ends and quantization to whole CASINO blocks then force a large overshoot, giving $P_{\mathrm{rep}} = 5.23$. Calibration remains consistent with the Ne fixed-length control. This demonstrates why CASINO block ends should be regarded simply as the points at which CASINO performs the check, and why the checks should be frequent compared with the expected stopping length.

VIII. SCOPE, LIMITATIONS, AND PORTABILITY

The validated statement is deliberately narrow: for the DMC total-energy estimator tested here, automatic stopping using the guarded rule produces no statistically resolved additional loss of error-bar calibration relative to matched fixed-length calculations using the same on-the-fly error-bar estimator. This is not a claim that the underlying DMC uncertainty estimator has exact nominal Gaussian coverage. Some fixed-length controls themselves give less than the nominal coverage, particularly at long correlation times. Nor is it a claim about systematic QMC errors.

Only the total energy has been tested. The same statistical logic may apply to another scalar estimator if the calculation can provide a reliable on-the-fly correlation-corrected error bar and a way to judge when that estimate is trustworthy, but this requires separate validation. All of the tests use real trial wave functions in the fixed-node approximation. Fixed-phase calculations with inherently complex trial wave functions, whose zeros form a set of lower dimension than a nodal surface, so that the singularities of the local energy that produce the heavy tails of Sec. II are differently distributed, have not been examined; the rule itself makes no assumption about the origin of the series, but its behavior there is untested.

For portability, another code needs, at each check, the accumulated number of moves or samples N, the current scalar mean, a correlation-corrected error bar $\hat{\sigma}$, an estimate of the uncertainty in that error bar, an estimate of the correlation time or effective sample size, and a diagnostic that the error-bar estimate is trustworthy rather than truncated by insufficient data. CASINO obtains these from on-the-fly reblocking; another implementation could use batch means or a spectral estimator. The checks need not coincide with output blocks, but they should be frequent enough that block or checkpoint spacing does not dominate the confirmation window.

The real-DMC validation uses a total target weight of 200 in the reference calculation and a small MPI decomposition. Because the statistical time series is formed from the globally reduced population mean, splitting a fixed total population over a different number of ranks does not change the estimator in principle. CASINO's present use of CPU-time-defined CASINO blocks does, however, make the spacing between checks depend on throughput and therefore indirectly on process count. A total target weight of 200 is modest, and for an all-electron calculation a population-control bias[11,12] of the DMC energies themselves cannot be excluded at this pop-

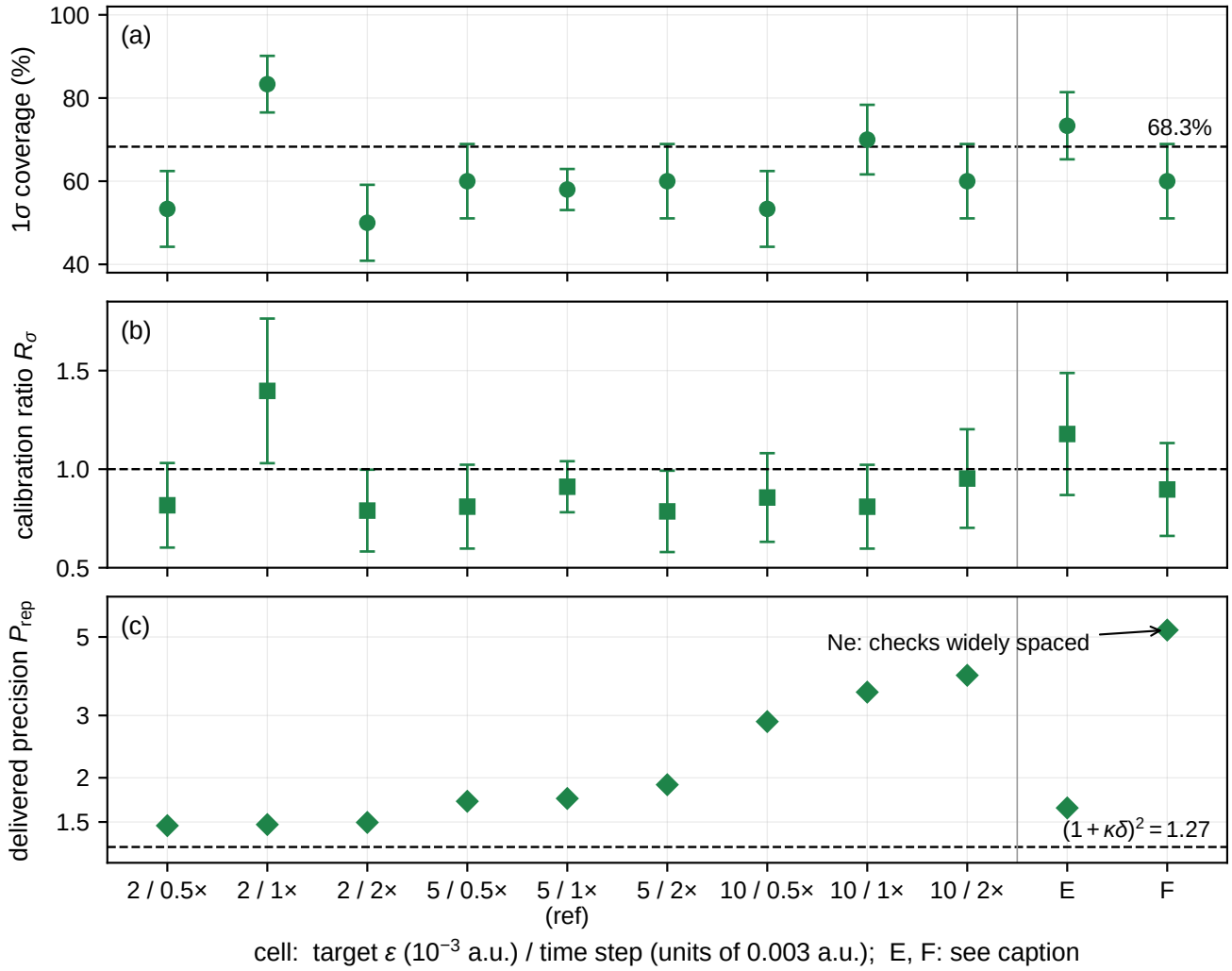


FIG. 4. Robustness of the guarded criterion with all rule parameters fixed. The cells are labeled on the horizontal axis as in Table II: the first number is the requested standard error ε in units of 10^{-3} a.u. and the second the DMC time step in units of the reference value 0.003 a.u.; "ref" marks the reference calculation, E is the degraded-wave-function O_2 test, and F the all-electron Ne test. (a) One-standard-error coverage of each guarded cell; error bars are binomial and the dashed line is the nominal Gaussian value. (b) Calibration ratio R_σ; the error bars are two standard errors, $2R_\sigma/\sqrt{2(n-1)}$ for a cell of n runs, from the sampling uncertainty of SD(E), and the dashed line is unity. (c) Delivered-precision factor $P_{\rm rep} = (\sigma_{\rm rep}/\varepsilon)^{-2}$, the median over the runs of each cell; the dashed line is the idealized margin-only boundary value $(1+\kappa\delta)^2 \simeq 1.27$ for $\delta \simeq 0.13$ discussed in Sec. VI. Fixed-length controls are used for the cell-by-cell comparisons reported in the text.

ulation. Such a bias does not enter the present conclusions. It is a property of the stationary population dynamics, the same for an automatically stopped run and for a fixed-length run made with the same population, time step, and trial wave function; since every automatic ensemble is scored against a reference energy obtained from fixed-length controls of exactly that kind, the bias is common to both and cancels from the coverage, the standardized errors, and the calibration ratio. What these tests cannot show is how the calibration of the error bar itself changes with population. Walker-population dependence has not been mapped systematically here; changing the population can alter population-control effects, genealogical correlations, and the variance of the population-averaged estimator, and is a natural target for further stress testing.

Larger systems. The systems tested here have at most 16 electrons, whereas automatic stopping would be of most value in expensive calculations on large molecules and periodic supercells. What follows is an expectation, not a validated result. Nothing in the rule refers to the size of the system: the series of Eq. (1) is a population average whatever the number of electrons, and every threshold ($\delta_{\rm max}$, the factor of 100 in Eq. (14), κ, and g) is

TABLE II. Per-cell results for the robustness grid. A cell is labeled by its requested standard error ε (in units of 10^{-3} a.u.) and its DMC time step δt (in units of the reference value 0.003 a.u.); E and F are defined in the text, and n is the number of guarded runs. "Calibration" is R_σ of Eq. (10); $\sigma_{\rm rep}/\varepsilon$ is the mean over the runs of the cell, and $P_{\rm rep}$ the median over those runs of the delivered-precision factor $(\sigma_{\rm rep}/\varepsilon)^{-2}$. The $n = 30$ calibration ratios have large sampling uncertainty and are interpreted as stress tests rather than precision measurements. For the 30/30 two-standard-error coverage entry, the exact two-sided 95% Clopper–Pearson interval is 88.4–100%.

Cell ($\varepsilon \times 10^3$ / δt)	n	Calibration	1σ cov. (%)	2σ cov. (%)	$\sigma_{\rm rep}/\varepsilon$	$P_{\rm rep}$
2 / 0.5×	30	0.817	53.3 ± 9.1	90.0 ± 5.5	0.83	1.46
2 / 1×	30	1.397	83.3 ± 6.8	100	0.82	1.47
2 / 2×	30	0.790	50.0 ± 9.1	90.0 ± 5.5	0.82	1.50
5 / 0.5×	30	0.810	60.0 ± 8.9	86.7 ± 6.2	0.77	1.72
5 / 1× (ref)	100	0.911	58.0 ± 4.9	95.0 ± 2.2	0.76	1.75
5 / 2×	30	0.786	60.0 ± 8.9	83.3 ± 6.8	0.73	1.91
10 / 0.5×	30	0.856	53.3 ± 9.1	93.3 ± 4.6	0.60	2.88
10 / 1×	30	0.810	70.0 ± 8.4	86.7 ± 6.2	0.53	3.49
10 / 2×	30	0.952	60.0 ± 8.9	96.7 ± 3.3	0.50	3.90
E (u-only Jastrow)	30	1.178	73.3 ± 8.1	96.7 ± 3.3	0.79	1.64
F (all-electron Ne)	30	0.897	60.0 ± 8.9	93.3 ± 4.6	0.42	5.23

a dimensionless number applied to the correlation structure measured in the run itself. What grows with system size is the scale of what is measured: the variance of the local energy, and with it the number of moves needed for a given absolute target, together with the cost of each move, while the correlation time in moves depends on the time step and the trial wave function. The rule responds to all of these through $\hat{\sigma}$ and $N_{\rm corr}$ alone. In particular, the resolution condition of Eq. (12) is met after about 23 reblocking blocks, which for the present block selector is between roughly 160 and 320 $N_{\rm corr}$ moves once the correlation estimate has settled; the gate as observed opens at the first check after all of its conditions hold, so that the spacing of the checks adds to this (for Ne it is the first CASINO block end, at about 25,000 moves). In the time-step grid of Sec. VII, where $N_{\rm corr}$ is about 36, 70, and 136 moves at the three time steps, the median observed gate-opening lengths were 6600 to 6900, 13,100 to 14,800, and 24,300 to 25,100 moves. A calculation whose target would be met sooner than that runs on until the gate opens and delivers a smaller error bar than requested, as in the loose-target cells of Table II; for an expensive system the gate, rather than the target, may thus set the minimum useful run, and a target unreachable within the available time is reported as such by the safeguard of Sec. V. The overhead from the spacing of the checks should conversely matter least for large systems, in which few moves fit into a CASINO block, the opposite limit from Ne. Two issues outside the stopping rule grow with system size: long equilibration, discussed next, and the allocation of effort across many calculations, as in a twist-averaged supercell study, which lies outside the inner problem treated here.

The rule assumes that, after equilibration, the mean and variance are stationary during statistics accumulation. The coarse drift check is only an additional safeguard against large pre-existing drift and should not be interpreted as an equilibration detector. Automatic detection of the initial equilibration transient is a separate problem,[22] and a pressing one for unattended operation: in realistic systems the equilibration of the walker population can take far longer than in the small all-electron systems studied here, which were equilibrated for 1.5 to 4 a.u. of imaginary time, and a residual transient biases the mean without necessarily being visible in the error bar. A monitor of residual non-equilibration would enter the logic of Fig. 2 naturally, as a further condition that must be passed, like the validity gate, before the target is tested. We have not constructed or validated such a monitor here.

CASINO also implements a complementary budget-limited mode that stops at an allocated resource limit and can optionally terminate earlier when the remaining statistically resolvable improvement is negligible. Appendix A gives the compact definition and validation summary. A purely clock-triggered stop does not select on $\hat{\sigma}$; an elective early budget stop does and therefore requires its own guarded confirmation. The detailed development and validation record is retained with the data archive rather than in the main narrative.

IX. CONCLUSIONS

Stopping a DMC calculation the first time its running uncertainty estimate falls below a target is unsafe for two distinct reasons. Early in the run, the correlation-corrected uncertainty estimate may not yet have resolved the correlation time at all. Later, even a mature uncertainty estimate is biased downward by first-crossing selection.

Both effects are visible directly in production DMC ensembles. In the reference O_2 calculation, 30.5% of naive runs stop during the immature-correlation regime, in extreme cases reporting uncertainties wrong by three orders of magnitude. The surviving late stops remain undercalibrated relative to fixed-length calculations.

A reliable automatic rule can be built from simple ingredients: require the validity gate to open, apply a safety margin to the target, and require the condition to remain satisfied while a substantial fraction of new data is accumulated. In the present implementation a 30% sample-growth window removes the detectable stopping-induced calibration defect. A matched-length fixed-run control gives essentially the same calibration ratio, coverage, and standardized-error width as the automatically stopped ensemble.

For the reference calculation the median cost is about 1.5 times that of an ideal fixed-length calculation whose required length is known in advance. The overhead is not universal and can be larger if the checks are too widely spaced; that dependence is an implementation issue rather than a requirement of the statistical method.

The method is implemented in CASINO, but its portable content is straightforward: another DMC code needs an on-the-fly correlation-corrected error bar, a way to determine when that error bar is trustworthy, and checks sufficiently frequent compared with the desired confirmation window. The central result is therefore not a new uncertainty estimator, but a way of using an existing one without introducing an additional bias simply because the calculation is allowed to decide for itself when to stop.

ACKNOWLEDGMENTS

The automatic-stopping capability was first suggested by N. D. Drummond and P. López Ríos in 2014, and this work exists because they asked for it. N. D. Drummond's comments on an earlier draft materially improved the framing and accessibility of the manuscript.

AUTHOR DECLARATIONS

Conflict of Interest

The author has no conflicts to disclose.

Author Contributions

M. D. Towler: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Visualization; Writing - original draft; Writing - review and editing.

DATA AVAILABILITY

The target-error implementation is available in CASINO from v3.0, and the budget-limited extension from v3.1.22. The computational record supporting this work, including validation specifications, per-run input decks and outputs, machine-readable stopping tables, analysis scripts, the trial-wave-function files and run records on which Appendix B is based, and documentation of the provenance and limitations of the synthetic tests, is available at Zenodo, DOI: 10.5281/zenodo.22281557. The CASINO source code and user documentation are available from Ref. 10.

Appendix A: Budget-limited (`small error`) stopping

A fixed resource budget defines a complementary problem. If a run could continue indefinitely then $\sigma \propto N^{-1/2}$ keeps decreasing, so there is no natural statistical stopping point corresponding to "as small an error bar as possible". Let $N_{\max}$ denote the projected *total* sample size affordable within the budget: the accumulated N plus what the remaining budget buys at the current sampling rate. Under the stabilized $N^{-1/2}$ scaling, the improvement still available at sample size N is estimated as

$$\Delta\hat{\sigma} = \hat{\sigma}(N)\left(1 - \sqrt{N/N_{\max}}\right). \tag{A1}$$

The optional diminishing-returns criterion, reported by CASINO as `RETURNS EXHAUSTED`, is

$$\Delta\hat{\sigma} < \kappa_s \,\mathrm{se}(\hat{\sigma}), \qquad \kappa_s = 1, \tag{A2}$$

so the run may end when the improvement still purchasable is smaller than the resolution with which the current uncertainty is known. Because Eq. (A2) itself selects on $\hat{\sigma}$, it is evaluated only after the same validity gate as the target-error rule and must persist over a confirmation window scaled to the remaining budget. A coarse drift check is retained. The run can then end in three normal ways: acquisition of an optional target-error floor using the full rule of Sec. IV; a guarded diminishing-returns stop satisfying Eq. (A2); or exhaustion of the resource budget. In CASINO these correspond to the banners `SMALL ERROR: FLOOR ACQUIRED`, `SMALL ERROR: RETURNS EXHAUSTED`, and `SMALL ERROR: BUDGET EXHAUSTED`, respectively. A separate user-specified hard cap on the number of DMC moves can bind before any of these and is reported as a fourth, explicitly distinguished outcome. At a pure budget exhaustion the clock, not $\hat{\sigma}$, triggers the stop; if the validity gate has never opened, the code stops without claiming a calibrated statistical uncertainty. As for the target-error rule, the budget applies to the current run segment only: after a restart the accumulated statistical state is restored but the CPU-time budget starts afresh.

The budget extension was validated in 650 production runs, including matched fixed-length controls for both

TABLE III. The trial wave functions of Appendix B. E_{VMC} is the variational energy and σ^2_{VMC} the variance of the local energy, from single VMC runs; E_{DMC} is the reference energy of the corresponding fixed-length control ensemble of Sec. VII, at $\delta t = 0.003$ a.u. for O_2 and 0.002 a.u. for Ne. Energies in a.u., variances in a.u.^{2}; parentheses give the uncertainty in the last digits.

Trial function	E_{VMC}	σ^2_{VMC}	E_{DMC}
O_2, reference	−149.996(12)	2.62	−150.27931(39)
O_2, degraded (E)	−149.843(19)	5.66	−150.27893(39)
Ne (F)	−128.852(23)	6.48	−128.92257(19)

clock-triggered and statistics-triggered endings. Relative to those controls, the one- and two-standard-error coverage differences were $+6.0 \pm 5.6$ and -2.7 ± 3.1 percentage points for the clock-triggered stops, and $+0.7 \pm 5.5$ and 0.0 ± 3.1 points for the diminishing-returns stops; the corresponding ratios of automatic to control R_σ were 1.022 and 0.954. No additional calibration defect was resolved. The present implementation also contains a slope check intended to suppress an early stop while the variance estimate is still changing anomalously. Successive points in this fit are cumulative and therefore share most of their data, so the usual least-squares error bar is too small. The measured rate of spurious triggers is therefore treated only as an empirical implementation diagnostic affecting efficiency, not as a confidence statement. The detailed derivation, failed first specification, two efficiency regimes, and complete per-run validation record are retained in the public data archive.

Appendix B: Trial wave functions

The three trial wave functions used in this work are of Slater-Jastrow form, $\Psi_T = e^J \sum_k c_k D_k$, with the Jastrow factor J of Ref. 23. Table III summarizes them and gives their variational energies. The Jastrow parameter files, the input files, and the outputs of every run quoted here are part of the data archive, which also identifies the two orbital files, both distributed with CASINO.

O_2, reference. The molecule is in its triplet ground state, with nine spin-up and seven spin-down electrons, at an internuclear distance of 2.283 a.u. The orbitals are numerical orbitals tabulated on the two-dimensional grid of the finite-difference Hartree-Fock program 2DHF for diatomic molecules,[24] and the determinantal part is an expansion of 12 determinants in these orbitals, dominated by a reference determinant of unit-magnitude coefficient; the other eleven coefficients, of which seven are independent, have magnitudes between 0.005 and 0.10. The Jastrow factor comprises an isotropic electron-electron term u, an electron-nucleus term χ, and an electron-electron-nucleus term f, together with two terms of cylindrical symmetry about the molecular axis, a two-body term u_{cyl} and a one-body term χ_{cyl}. Orbitals, determinant coefficients, and Jastrow factor are those of the O_2 test case distributed with CASINO; none was optimized for this work.

O_2, degraded (cell E). The orbitals and the determinant expansion are those of the reference function. The Jastrow factor is reduced to the u term alone (expansion order 8, spin dependent), whose parameters and cutoff length were reoptimized by variance minimization in two cycles of 10,000 configurations with the determinant coefficients held fixed. The u term imposes the electron-electron cusp conditions, and the electron-nucleus cusps are a property of the numerical orbitals, so no cusp condition is violated; what is removed is the explicit electron-nucleus and electron-electron-nucleus correlation. The variational energy rises by 0.153(23) a.u. and the variance of the local energy slightly more than doubles. Because the determinantal part is unchanged, the two O_2 functions have the same nodal surface, and their fixed-node DMC energies at the reference time step agree within statistical error (Table III).

Ne (cell F). The orbitals are restricted Hartree-Fock orbitals expanded in a Gaussian basis of s and p functions (13 shells, 25 basis functions), with the electron-nucleus cusp correction of CASINO applied,[9] in a single determinant. The Jastrow factor has u and f terms, whose parameters were optimized for this work in three cycles of 10,000 configurations by minimizing the mean absolute deviation of the local energies from their median.

The variational energies of Table III are from single VMC runs, of 10^5 steps for O_2 and 4×10^4 steps for Ne, with error bars from on-the-fly reblocking; they characterize the trial functions and are not converged benchmarks. Repeat runs in the archive agree with them within the quoted errors.

USE OF ARTIFICIAL INTELLIGENCE

Generative AI tools were used in this work: Claude (Anthropic; Opus 4.1 and Fable 5.1) and ChatGPT (OpenAI; GPT-5 and GPT-6). They were used to write and check analysis scripts, to extract and tabulate quantities from the archived run outputs, to check the manuscript against those outputs, and to draft and revise text, so that the analysis and its checking could be done faster and more thoroughly. All calculations were performed with CASINO, every number reported was checked against the archived data, and the author takes full responsibility for the content.

[1] A. Wald, *Sequential Analysis* (Wiley, New York, 1947).
[2] Y. S. Chow and H. Robbins, Ann. Math. Statist. **36**, 457 (1965).
[3] P. W. Glynn and W. Whitt, Ann. Appl. Probab. **2**, 180 (1992).
[4] J. M. Flegal, M. Haran, and G. L. Jones, Statist. Sci. **23**, 250 (2008).
[5] D. Vats, J. M. Flegal, and G. L. Jones, Biometrika **106**, 321 (2019).
[6] G. L. Jones, M. Haran, B. S. Caffo, and R. Neath, J. Am. Stat. Assoc. **101**, 1537 (2006).
[7] J. M. Flegal and L. Gong, Stat. Sinica **25**, 655 (2015).

[8] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon, Ann. Stat. **49**, 1055 (2021).
[9] R. J. Needs, M. D. Towler, N. D. Drummond, P. López Ríos, and J. R. Trail, J. Chem. Phys. **152**, 154106 (2020).
[10] CASINO project website and documentation, https://vallico.net/casinoqmc.
[11] C. J. Umrigar, M. P. Nightingale, and K. J. Runge, J. Chem. Phys. **99**, 2865 (1993).
[12] N. Cerf and O. C. Martin, Phys. Rev. E **51**, 3679 (1995).
[13] H. Flyvbjerg and H. G. Petersen, J. Chem. Phys. **91**, 461 (1989).
[14] D. R. Kent IV, R. P. Muller, A. G. Anderson, W. A. Goddard III, and M. T. Feldmann, J. Comput. Chem. **28**, 2309 (2007).
[15] R. M. Lee, G. J. Conduit, N. Nemec, P. López Ríos, and N. D. Drummond, Phys. Rev. E **83**, 066706 (2011); **85**, 029908(E) (2012).
[16] U. Wolff, Comput. Phys. Commun. **156**, 143 (2004); erratum **176**, 383 (2007).
[17] M. Jonsson, Phys. Rev. E **98**, 043304 (2018).
[18] V. R. Pandharipande, S. C. Pieper, and R. B. Wiringa, Phys. Rev. B **34**, 4571 (1986).
[19] J. R. Trail, Phys. Rev. E **77**, 016703 (2008).
[20] J. R. Trail, Phys. Rev. E **77**, 016704 (2008).
[21] P. López Ríos and G. J. Conduit, Phys. Rev. E **99**, 063312 (2019).
[22] J. D. Chodera, J. Chem. Theory Comput. **12**, 1799 (2016).
[23] N. D. Drummond, M. D. Towler, and R. J. Needs, Phys. Rev. B **70**, 235119 (2004).
[24] J. Kobus, L. Laaksonen, and D. Sundholm, Comput. Phys. Commun. **98**, 346 (1996).