The AutoCorrelation Integral Drill (ACID) Test Set

Gözdenur Toraman,[†] Dieter Fauconnier,[†‡] and Toon Verstraelen[*¶]

† Soete Laboratory, Ghent University, Technologiepark-Zwijnaarde 46, 9052 Ghent, Belgium
‡ FlandersMake@UGent, Core Lab MIRO, 3001 Leuven, Belgium
¶ Center for Molecular Modeling (CMM), Ghent University, Technologiepark-Zwijnaarde 46, B-9052, Ghent, Belgium

*E-mail: toon.verstraelen@ugent.be

Version v1.2.1 (2026-05-04 9b7c6de)

Summary

The data set consists of synthetic time-correlated sequences of varying lengths, generated using different covariance kernels.

The purpose of the data set is to validate algorithms for estimating the integral of an autocorrelation function, which is relevant for uncertainty quantification and the estimation of transport properties. The first application was to validate the algorithm implemented in STACIE.

The set contains in total 15360 test cases, and each case consists of one or more time series. They are organized such that one can systematically study the convergence of the statistical estimate of the autocorrelation integral (and its uncertainty) with increasing sequence length (N) and increasing number of sequences used as input (M).

License

Overview of the data

Covariance kernels are constructed with one or two of the following three models. In all models, the parameter A_0 corresponds to the integral of the autocorrelation function for that specific contribution. The three kernel models in continuous time and frequency domains are described by their autocorrelation function (ACF):

$$c(\Delta_t) = \mathrm{COV}[\hat{x}(t), \hat{x}(t + \Delta_t)]$$

or equivalently their power spectral distribution (PSD):

$$C(f) = \int_{-\infty}^{\infty} c(\Delta_t) e^{-2\pi i f \Delta_t} \, \mathrm{d}\Delta_t$$

1. The **white noise** model consists of uncorrelated data and has the following ACF:

$$c(\Delta_t) = A_0 \delta(\Delta_t)$$

The PSD is constant:

$$C(f) = A_0$$

This model will be denoted as $\mathrm{W}(A_0)$.

2. The **exponential model** has an exponentially decaying ACF:

$$c(\Delta_t) = \frac{A_0}{2\tau} \exp\left(-\frac{|\Delta_t|}{\tau}\right)$$

where τ is the exponential autocorrelation time and A_0 is the integral of the autocorrelation function. The PSD is:

$$C(f) = \frac{A_0}{1 + (2\pi f \tau)^2}$$

This model will be denoted as $\mathrm{E}(A_0, \tau)$.

3. The **stochastic harmonic oscillator** was adapted from the work of Foreman-Mackey et al. It's ACF (with modified normalization conventions) is:

$$c(\Delta_t) = A_0 \pi f_0 Q \exp\left(-\frac{\pi f_0 \, |\Delta_t|}{Q}\right) \begin{cases} \cosh(\eta 2\pi f_0 \Delta_t) + \frac{1}{2\eta Q} \sinh(\eta 2\pi f_0 \, |\Delta_t|) & \text{if} \quad 0 < Q < \frac{1}{2} \\ 1 + 2\pi f_0 \, |\Delta_t| & \text{if} \quad Q = \frac{1}{2} \\ \cos(\eta 2\pi f_0 \Delta_t) + \frac{1}{2\eta Q} \sin(\eta 2\pi f_0 \, |\Delta_t|) & \text{if} \quad Q > \frac{1}{2} \end{cases}$$

with

$$\eta = \left| \frac{1}{4Q^2} - 1 \right|^{\frac{1}{2}}$$

The PSD is:

$$C(f) = \frac{A_0 f_0^4}{\left(f^2 - f_0^2\right)^2 + \left(f f_0 / Q\right)^2}$$

where Q represents the quality of the oscillator, f_0 is the angular resonant frequency, and A_0 is the zero-frequency limit of the spectrum. (Note that Foreman-Mackeyet al. use a parameter $S_0 = \frac{A_0}{2}$, a unitary normalization convention for the Fourier transform and an angular frequency. These differences are only a matter of notation.)

This model will be denoted as $\mathrm{S}(A_0, f_0, Q)$

Using these three models, 12 covariance kernels are defined in Table 1 and were used to generated time-correlated sequences.

Kernel	**Definition**	τ_{int}
exp1p	$\mathrm{E}(1.0, 5.0)$	5.000
exp1w	$\mathrm{E}(0.9, 5.0) + \mathrm{W}(0.1)$	2.632
exp2	$\mathrm{E}(0.5, 2.0) + \mathrm{E}(0.5, 5.0)$	2.857
sho1pcrit	$\mathrm{S}(1.0, 0.04, 0.5)$	7.958
sho1pover	$\mathrm{S}(1.0, 0.15, 0.2)$	5.305
sho1punder	$\mathrm{S}(1.0, 0.03, 1.4)$	3.789
sho1wcrit	$\mathrm{S}(0.9, 0.04, 0.5) + \mathrm{W}(0.1)$	3.194
sho1wover	$\mathrm{S}(0.9, 0.15, 0.2) + \mathrm{W}(0.1)$	2.705
sho1wunder	$\mathrm{S}(0.9, 0.03, 1.4) + \mathrm{W}(0.1)$	2.286
sho2crit	$\mathrm{S}(0.8, 0.04, 0.5) + \mathrm{S}(0.2, 0.35, 0.1)$	6.920
sho2over	$\mathrm{S}(0.8, 0.15, 0.3) + \mathrm{S}(0.2, 0.35, 0.1)$	3.701
sho2under	$\mathrm{S}(0.8, 0.03, 1.4) + \mathrm{S}(0.2, 0.35, 0.1)$	3.920

Table 1: Summary of kernels used in the ACID test set.

For each kernel, sequences with $N = 1024$, 4096, 16384 and 65536 steps are generated, using a dimensionless time step $h = 1$. (In fact, sequences of double this length are generated with a discrete Fourier transform and the second half is discarded to obtain aperiodic sequences.) For each kernel and each number of steps, independent test cases are created comprising $M = 1$, 4, 16, 64, and 256 independent sequences. To ensure statistical robustness, 64 repetitions with unique random seeds are included for every combination of kernel, number of steps and number of sequences.

Example sequences, ACFs and PSDs for all kernels are shown in Figures Figure 1, Figure 2 and Figure 3, respectively.

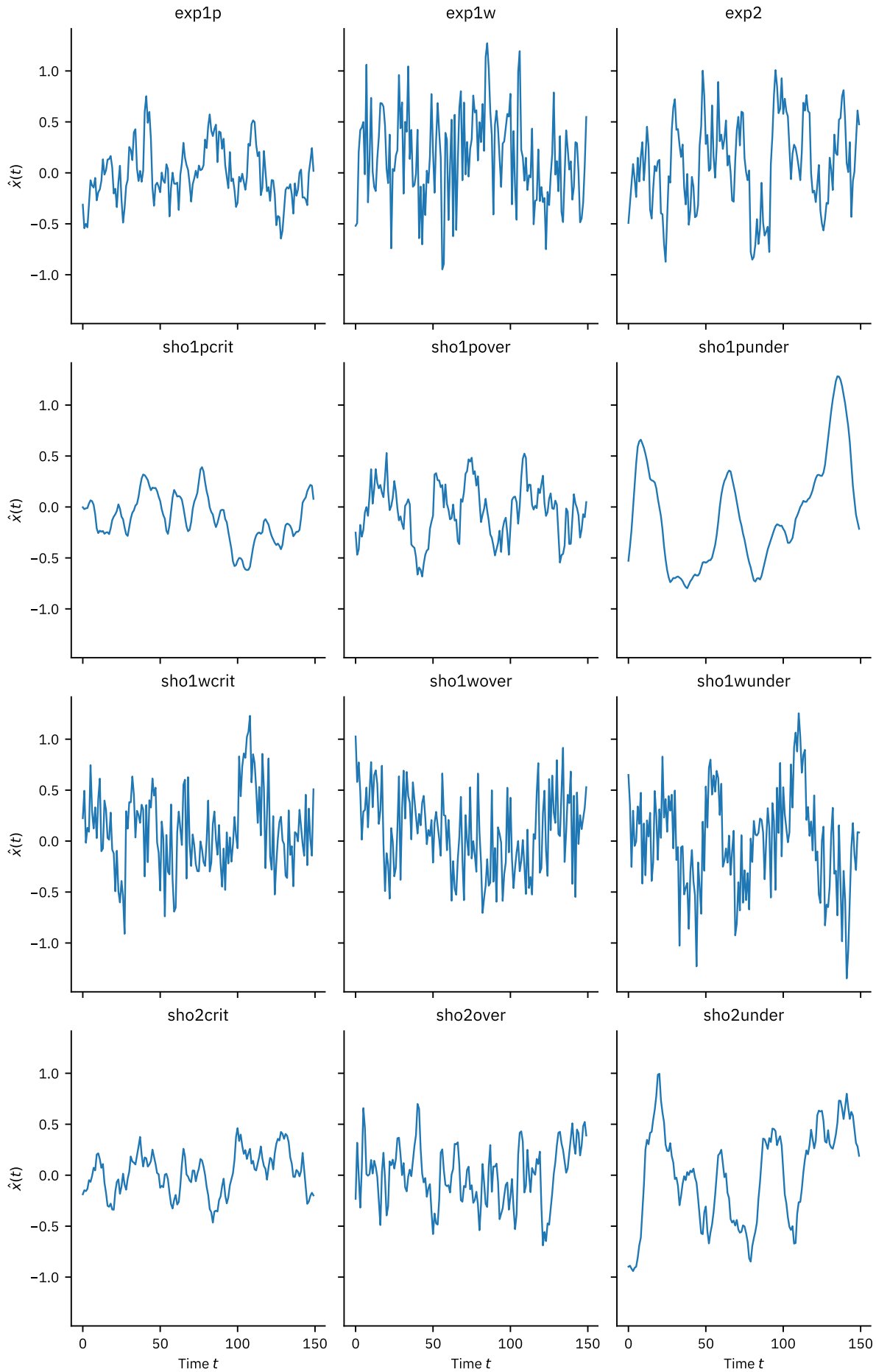

Figure 1: Example sequences obtained with each kernel. (First 150 steps of the first sequence in the first out of 64 test cases for $N = 1024$ and $M = 256$.)

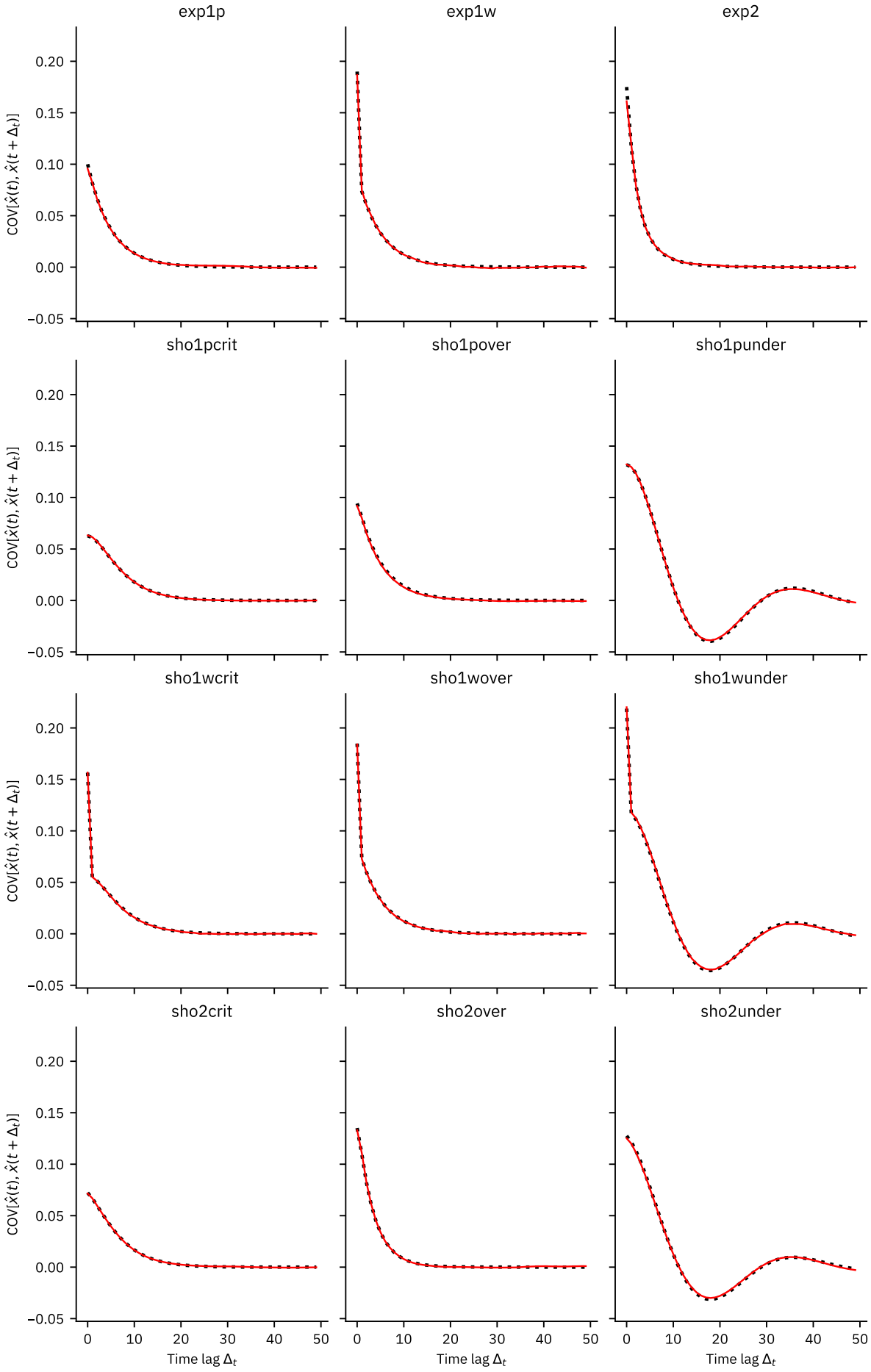

Figure 2: Autocorrelation functions of the kernels. The analytical model is plotted as a dotted black line. The empirical ACF derived from the first out of 64 test cases for $N = 1024$ and $M = 256$ is plotted as a red solid line.

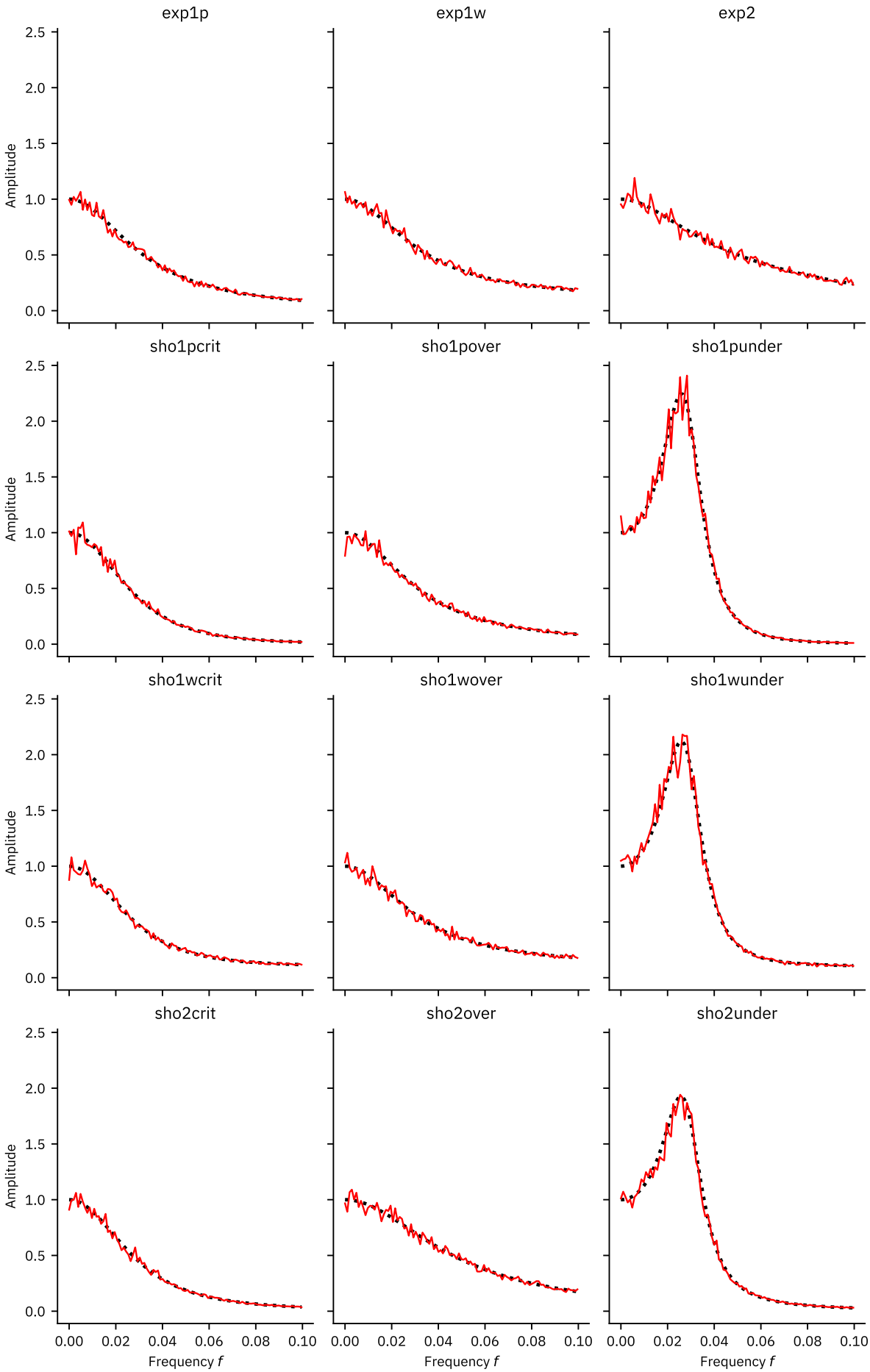

Figure 3: Power spectral distributions (PSDs) of the kernels. The analytical model is plotted as a dotted black line. The empirical PSD (periodogram) derived from the first out of 64 test cases for $N = 1024$ and $M = 256$ is plotted as a red solid line.

All kernels have an autocorrelation integral of 1. They are all parametrized to have an almost quadratic PSD close to zero frequency, with deviations less than 2.5% RMS for the first 20 grid points of the spectrum and less than 10% for the first 40 points. This has two important implications on the data:

- It guarantees that also the shortest synthetic sequences (1024 steps) are just long enough to capture the slowest time correlations. (For longer sequences, the deviation from the quadratic fit are much smaller.)
- For the spectra averaged over 256 sequences, the relative error is about $\frac{1}{\sqrt{256}}$, which corresponds to 6.25%. This is larger than the systematic deviation between the quadratic model and the real PSD for the first 20 points.

For each combination of kernel, sequence length and number of sequences, data are stored in ZARR version 3 ZIP archives, using the pattern `{kernel_name}_nstep{nstep:05d}_nseq{nseq:04d}.zip`. The data stored in each ZARR file are described in Table 2.

ZARR field	Description
`root.attrs["corrtime_int"]`	The integrated autocorrelation time
`root.attrs["typst"]`	A typst equation describing the kernel
`root.attrs["latex"]`	A latex equation describing the kernel
`root["times"]`	The time axis of the sequences
`root["freqs"]`	The frequency axis of the power spectrum
`root["omegas"]`	$2\pi \times$ the frequency axis
`root["psd"]`	The reference power spectrum with normalization conventions given above
`root["acf"]`	The reference autocorrelation function
`root["sequences"]`	The stochastic time-dependent sequences

Table 2: Overview of data stored in each ZARR file.

All arrays, except `sequences` are 1D arrays. The sequences are stored in a 3D array with shape `(ncase, nseq, nstep)`, where `ncase` is 64, `nseq` is the number of sequences (M) and `nstep` is the number of steps (N). The ground truth of the autocorrelation integral is `root["psd"][0]`.

Data generation

All Python scripts required for data generation and analysis are included in the archive. These scripts make use of open-source software libraries (see below).

The script `plan.py` defines the workflow to reconstruct the entire dataset from scratch. It can be executed with StepUp as follows on the command line:

```
stepup boot -n 8
```

where `8` is the number of parallel workers.

Remark on determinism

The data generation scripts are fully deterministic, meaning that running them multiple times on the same hardware and software environment will yield identical results. However, due to differences in floating-point arithmetic across different hardware architectures and software versions, running the StepUp workflow on different systems may lead to slight numerical differences in the generated data. (Even with identical data, the hashes of the ZIP files may differ due to operating system details

and software versions.) These differences should not affect the overall validity of the dataset, but they may impact the exact values derived from an analysis.

Software used

The following software is required to use the dataset:

- Python >= 3.12
- NumPy == 2
- Zarr == 3

To fully reconstruct the dataset, the following additional Python packages are required:

- StepUp >= 3.1.2
- StepUp RepRep >= 3.1.4