University of Oxford
Mathematical Institute

InFoMM
Industrially Focused
Mathematical Modelling

Stochastic Analysis and Correction of Floating Point Errors in Monte Carlo Simulations

Oliver Sheridan-Methven

27th February 2019

Supervisors:
Prof. Mike Giles — Oxford
Dr Christopher Goodyer — Arm

EPSRC
Engineering and Physical Sciences
Research Council

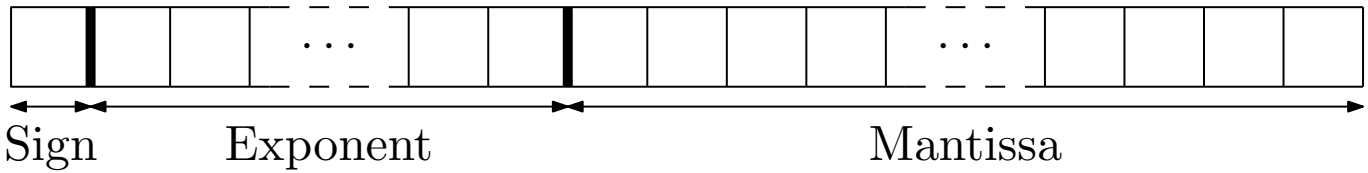
Overview

1. Fast computations and low-precision
2. Low-precision and higher speed
3. Are crude approximations useful?
4. Mixed-precision multilevel Monte Carlo
5. Convergence results
6. Recent results
7. Numerical results
8. Conclusions

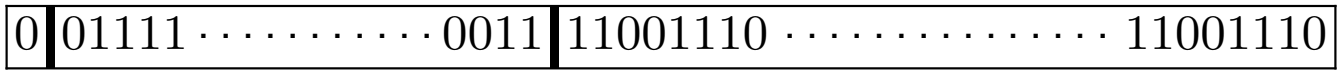
Reduced-precision

$$x \equiv 0.0001101011010101\ldots$$

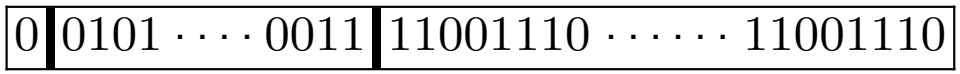
$$x \equiv \pm(1.a_1a_2\cdots a_m\cdots)_2 \times 2^{(b_1b_2\cdots b_e)_2}$$



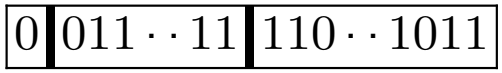
64-bit (Double)



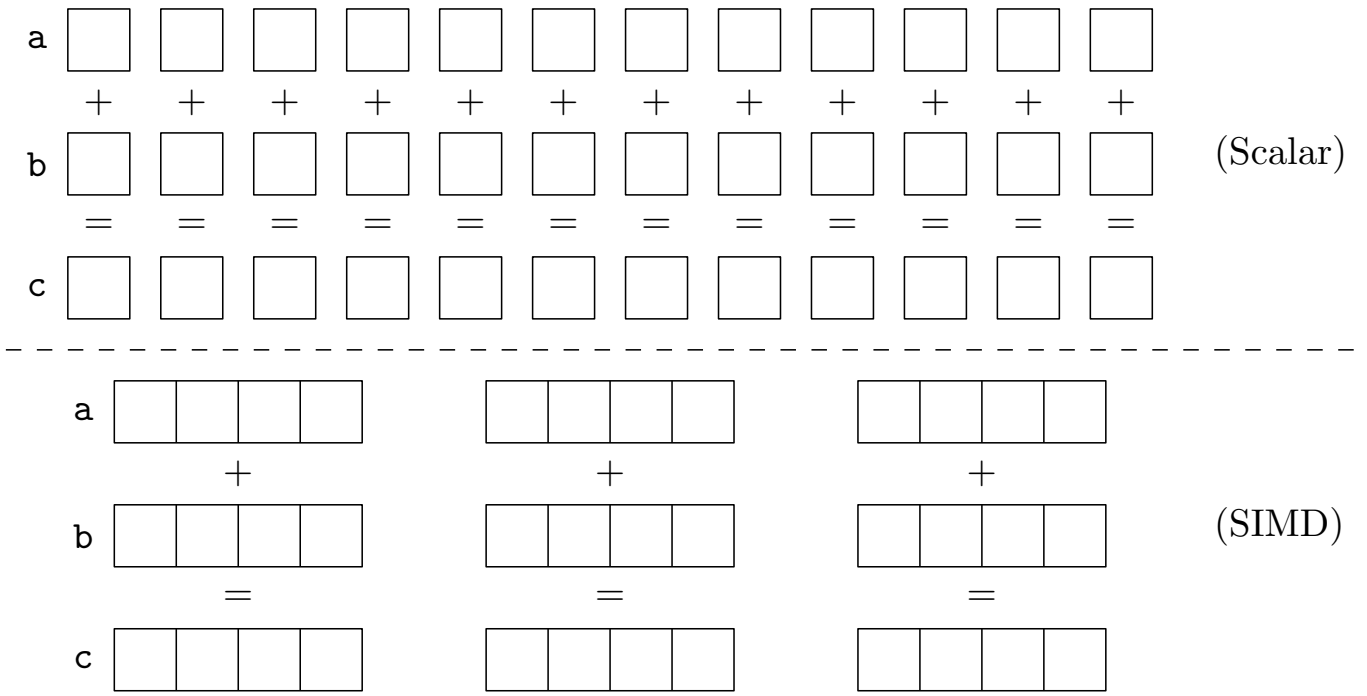
32-bit (Single)

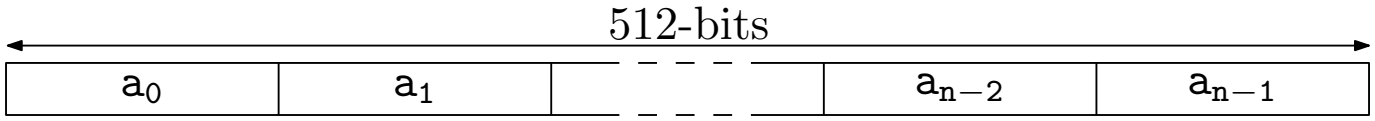


16-bit (Half)



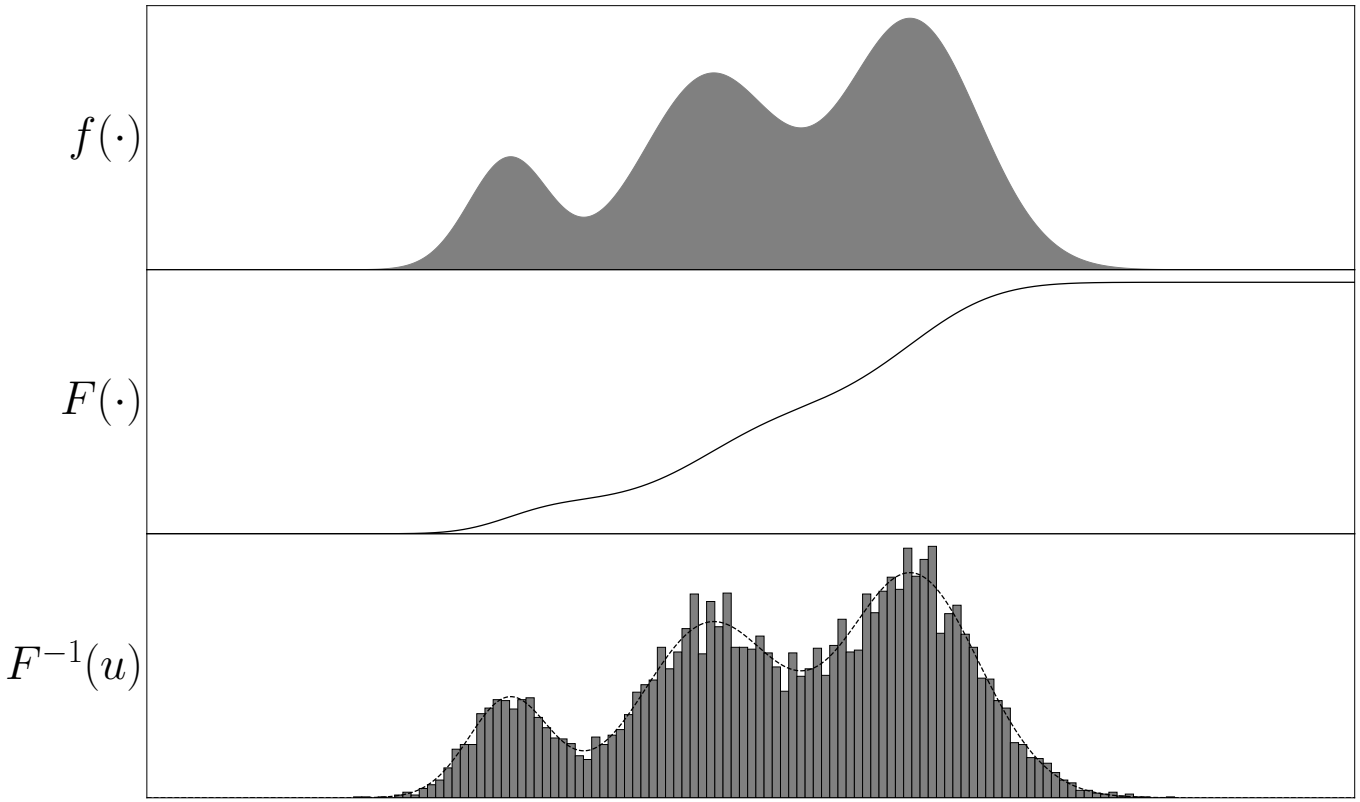
Vectorised (SIMD) operations



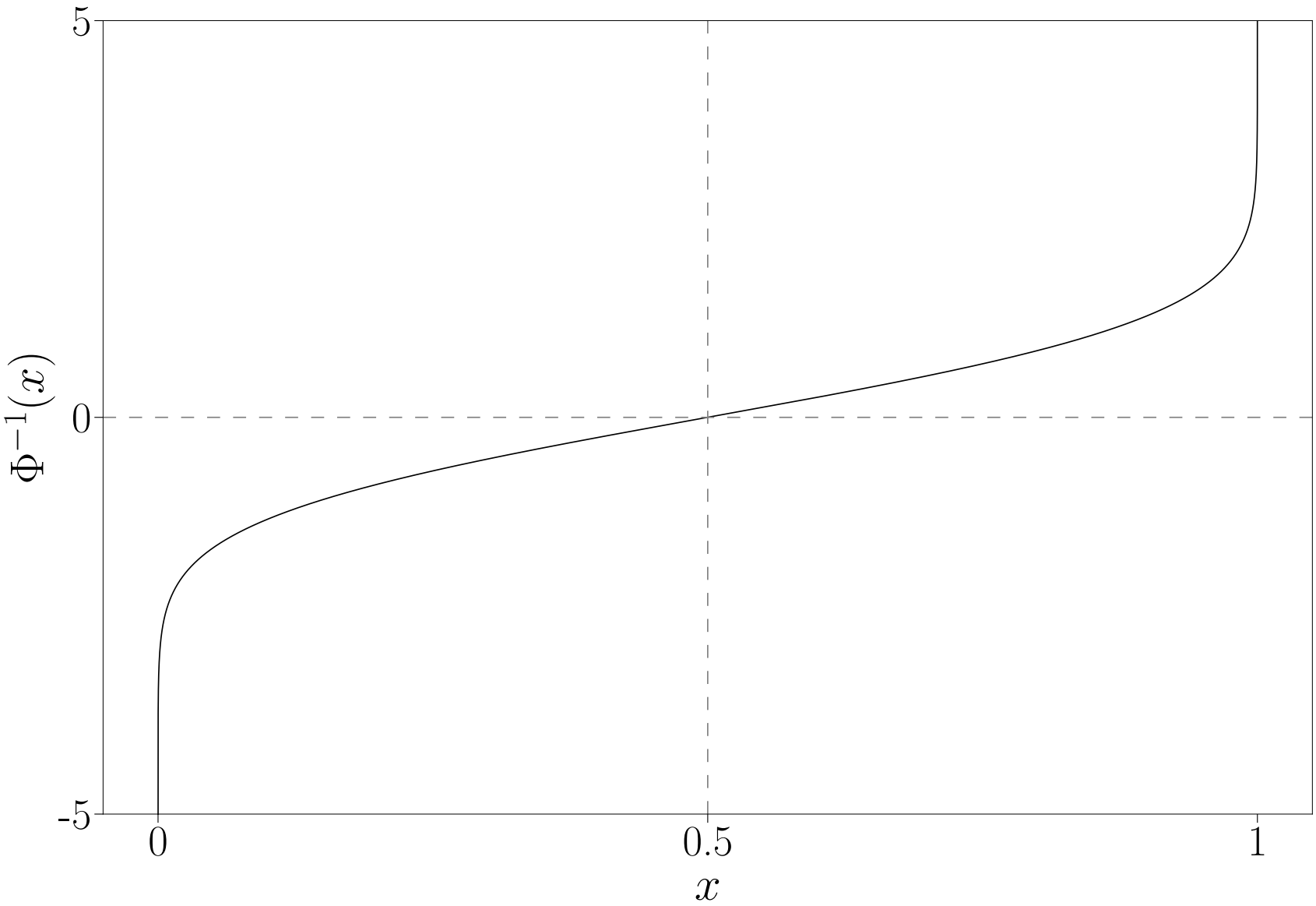


Mathematical situations of interest

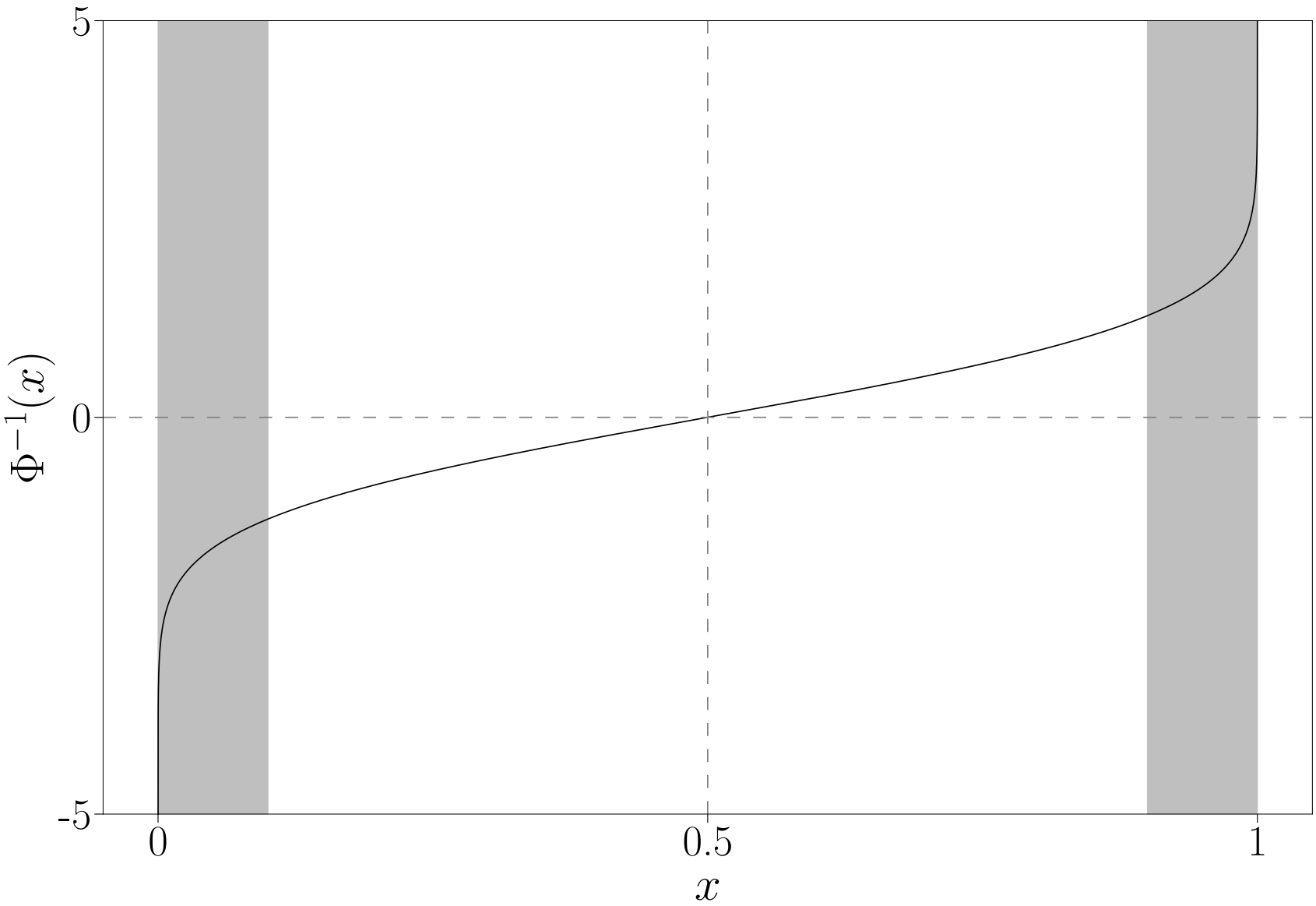
While there are many areas of mathematics which we could turn our attention to, we have been focusing on the numerical approximation of SDEs through Monte Carlo simulation. One of the primary areas where these concerns arise is in the simulation of random numbers from a distribution of interest, such as the Gaussian distribution.



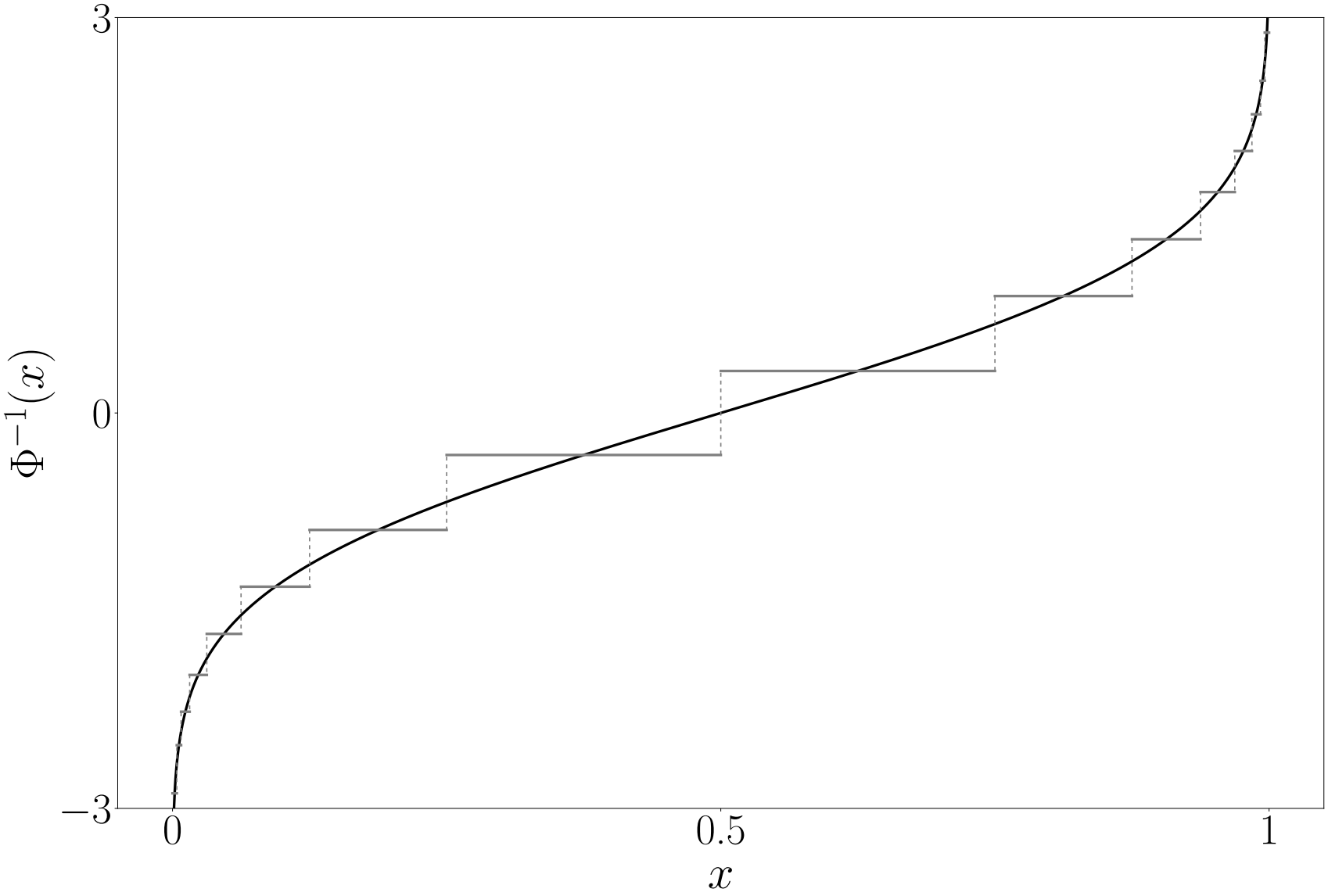
The inverse Gaussian CDF $\Phi^{-1}(\cdot)$



The inverse Gaussian CDF $\Phi^{-1}(\cdot)$



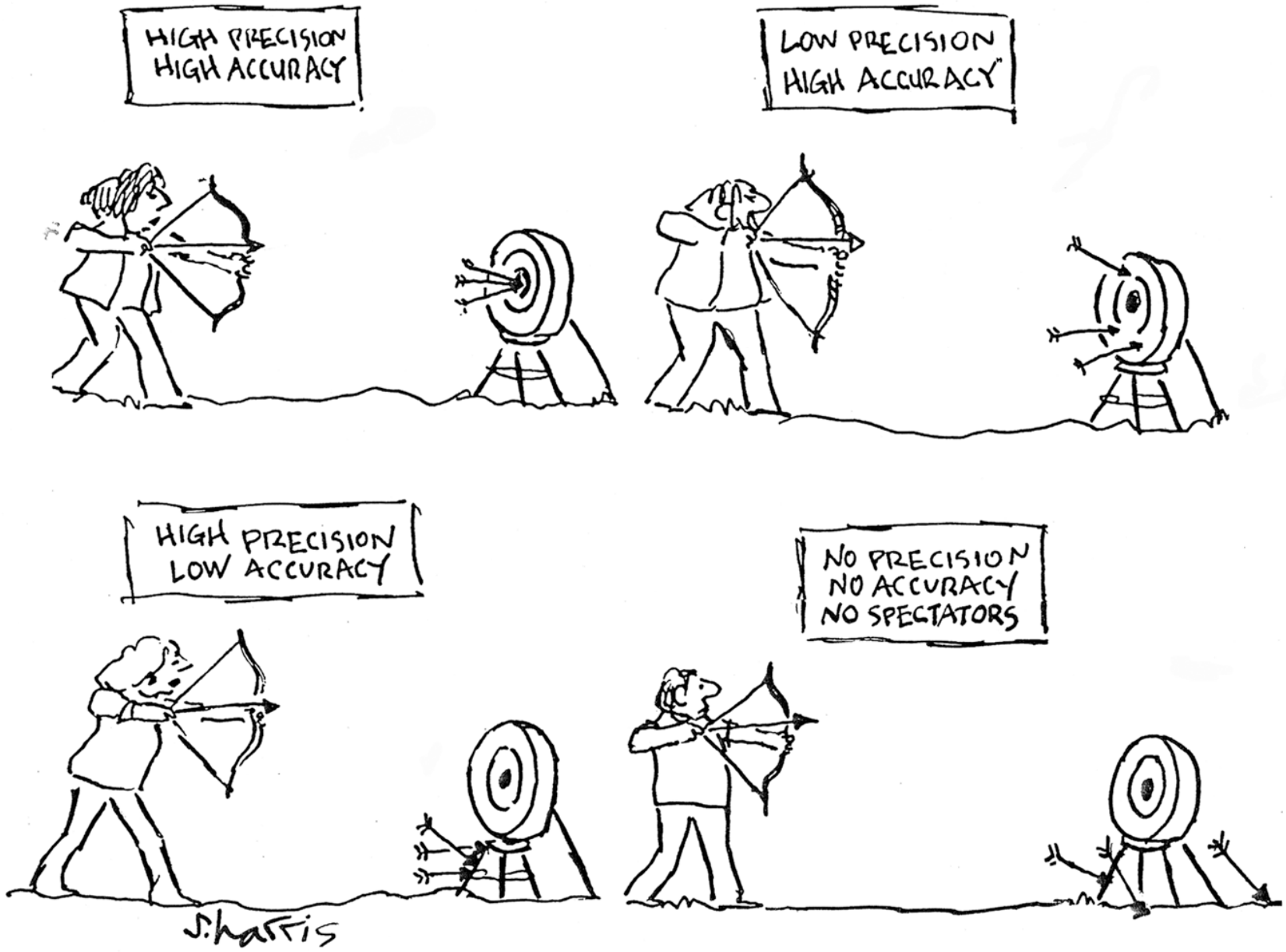
Piecewise constant approximation



Speed of execution

	Average speed (clock cycles)
Intel MKL	8
Lookup table	2

Accuracy and precision



$$\boxed{\mathbb{E}(P) \approx \mathbb{E}\big(\hat{P}_L\big) = \mathbb{E}\big(\hat{P}_0\big) + \sum_{l=1}^{L} \mathbb{E}\big(\hat{P}_l - \hat{P}_{l-1}\big).} \tag{1}$$

where $\mathbb{V}\big(\hat{P}_l - \hat{P}_{l-1}\big) \ll \mathbb{V}\big(\hat{P}_l\big)$ is achieved by **using the same underlying uniform random variables**.

Some possible multilevel constructions:

(Precision)

$$\mathbb{E}\big(\hat{P}_0\big) \to \mathbb{E}\big(\hat{P}_{\text{32-bit}}\big) + \mathbb{E}\big(\hat{P}_{\text{64-bit}} - \hat{P}_{\text{32-bit}}\big)$$
$$\mathbb{E}\big(\hat{P}_0\big) \to \mathbb{E}\big(\hat{P}_{\text{16-bit}}\big) + \mathbb{E}\big(\hat{P}_{\text{32-bit}} - \hat{P}_{\text{16-bit}}\big)$$

(Distribution)

$$\mathbb{E}\big(\hat{P}_0\big) \to \mathbb{E}\big(\hat{P}_{\text{LT}_{1024}}\big) + \mathbb{E}\big(\hat{P}_Z - \hat{P}_{\text{LT}_{1024}}\big)$$
$$\mathbb{E}\big(\hat{P}_0\big) \to \mathbb{E}\big(\hat{P}_{\text{cubic}}\big) + \mathbb{E}\big(\hat{P}_Z - \hat{P}_{\text{cubic}}\big)$$

We have the following SDE which we cannot solve:

$$\mathrm{d}X_t = a(t, X_t)\,\mathrm{d}t + b(t, X_t)\,\mathrm{d}W_t\,, \tag{2}$$

where we wish to estimate some property which depends on the solution at the final time

$$\mathbb{E}(P(X_T))$$

(or possibly the entire duration).

The Euler-Maruyama scheme

$$\hat{X}_{n+1} = \hat{X}_n + a(\tau_n, \hat{X}_n)\delta + b(\tau_n, \hat{X}_n)\sqrt{\delta}\, Z_n \tag{3}$$

The Euler-Maruyama scheme

$$\hat{X}_{n+1} = \hat{X}_n + a(\tau_n, \hat{X}_n)\delta + b(\tau_n, \hat{X}_n)\sqrt{\delta}\, Z_n \tag{3}$$

Approximate Euler-Maruyama schemes

$$\boxed{\hat{X}_{n+1} = \hat{X}_n + a(\tau_n, \hat{X}_n)\delta + b(\tau_n, \hat{X}_n)\sqrt{\delta}\, Z_n}$$

$$\tilde{X}_{n+1} = \tilde{X}_n + a(\tau_n, \tilde{X}_n)\delta + b(\tau_n, \tilde{X}_n)\sqrt{\delta}\, \tilde{Z}_n \tag{4}$$

$$\check{X}_{n+1} = \check{X}_n + a(\tau_n, \check{X}_n)\delta + b(\tau_n, \check{X}_n)\sqrt{\delta}\, Z_n + \epsilon_n \tag{5}$$

$$\bar{X}_{n+1} = \bar{X}_n \oplus \left(\bar{a}\left(\bar{\tau}_n, \bar{X}_n\right) \otimes \bar{\delta} \oplus \bar{b}\left(\bar{\tau}_n, \bar{X}_n\right) \otimes \left(\overline{\mathrm{sqrt}}\left(\bar{\delta}\right) \otimes \bar{Z}_n\right)\right) \tag{6}$$

Approximate Euler-Maruyama schemes

$$\boxed{\hat{X}_{n+1} = \hat{X}_n + a(\tau_n, \hat{X}_n)\delta + b(\tau_n, \hat{X}_n)\sqrt{\delta}\, Z_n}$$

$$\tilde{X}_{n+1} = \tilde{X}_n + a(\tau_n, \tilde{X}_n)\delta + b(\tau_n, \tilde{X}_n)\sqrt{\delta}\, \tilde{Z}_n \tag{4}$$

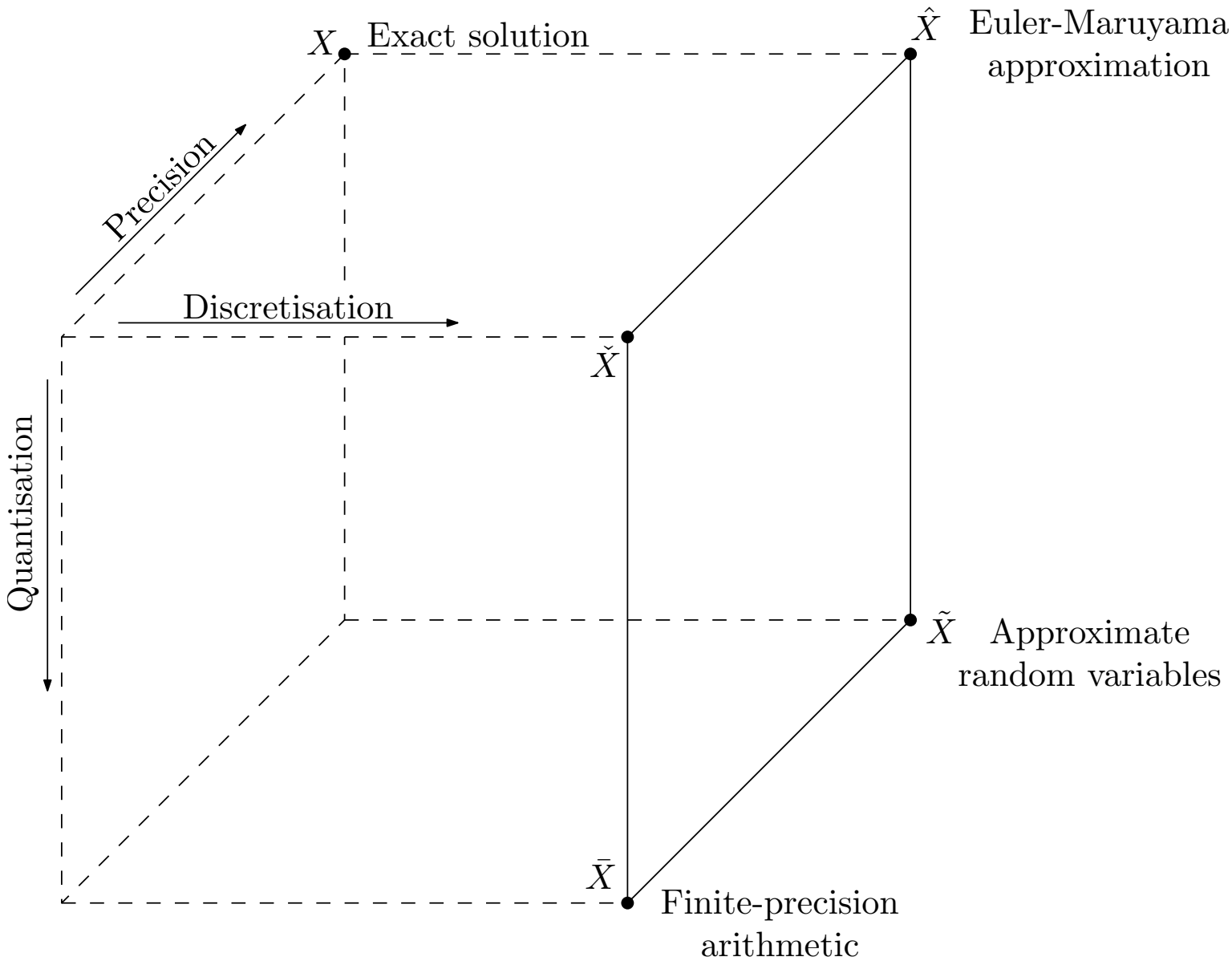
$$\check{X}_{n+1} = \check{X}_n + a(\tau_n, \check{X}_n)\delta + b(\tau_n, \check{X}_n)\sqrt{\delta}\, Z_n + \epsilon_n \tag{5}$$

(Arciniega and Allen [1])

$$\bar{X}_{n+1} = \bar{X}_n \oplus \left(\bar{a}\left(\bar{\tau}_n, \bar{X}_n\right) \otimes \bar{\delta} \oplus \bar{b}\left(\bar{\tau}_n, \bar{X}_n\right) \otimes \left(\overline{\mathrm{sqrt}}\left(\bar{\delta}\right) \otimes \bar{Z}_n\right)\right) \tag{6}$$

(Omland [2])

Types of approximation



Strong convergence I

Theorem

Under Assumptions 11.1 to 11.6 where the constants do not depend on δ, and assuming $\mathbb{E}\big(\tilde{Z}\big) = 0$ and $\mathbb{V}\big(\tilde{Z}\big) < \infty$, then there exists a constant $K > 0$, independent of N, δ, and q, such that

$$\mathbb{E}\left(\sup_{n \leq N} \left|\hat{X}_n - \tilde{X}_n\right|^2\right) \leq K\mathbb{E}\Big(\left|\tilde{Z} - Z\right|^2\Big). \tag{7}$$

Theorem

Under Assumptions 11.1 to 11.6 where the constants do not depend on δ, we assume there exists a symmetric approximate normal distribution $\tilde{Z}$ with a finite $(2p)$-th moment such that $\mathbb{E}(\tilde{Z}^{2p-1}) = 0$ and $\mathbb{E}(|\tilde{Z}|^{2p}) < \infty$ for a finite integer $1 < p < \infty$. There exists a finite constant $0 < K_p < \infty$, independent of N, δ, and q, but dependent on p, such that

$$\mathbb{E}\left(\sup_{n \leq N} \left|\hat{X}_n - \tilde{X}_n\right|^{2p}\right) \leq K_p\mathbb{E}\Big(\left|\tilde{Z} - Z\right|^{2p}\Big). \tag{8}$$

Theorem

[1, Theorem 2.2] Under Assumptions 11.1 to 11.6 where the constants do not depend on δ, we use the adapted Euler-Maruyama scheme in (5). With i.i.d. errors $\epsilon_n \sim \mathcal{N}$ with $\mathbb{E}(\epsilon) = 0$ and $\mathbb{V}(\epsilon) \leq C\varrho^2$, where C is a some finite and strictly positive constant and $\varrho \propto 2^{-r}$, then there exists a constant $K > 0$, independent of N and r, such that

$$\mathbb{E}\left(\sup_{n \leq N} \left|\hat{X}_n - \check{X}_n\right|^2\right)^{\frac{1}{2}} \leq K\sqrt{N}2^{-r}. \tag{9}$$

Strong convergence III

$$\mathscr{R}(x \circledast y) = (x * y)(1 + \varepsilon), \tag{10}$$

$$|\bar{a}(t, x) - a(t, x)| \leq C2^{-r}(1 + |a(t, x)|), \tag{11}$$

Theorem

For all precisions $r \geq \lceil \log_2(N + 3) + 1 \rceil$, under Assumptions 11.1 to 11.4, assuming the roundoff model (10) and approximation model (11), and if the time domain $[t_0, T]$ is normalised to $[0, 1]$, and the times are partitioned on dyadic intervals such that $\delta = 2^{-k}$ for some $k \in \mathbb{Z}$, then the strong L^2-error converges as

$$\mathbb{E}\left(\sup_{n \leq N} \left| \tilde{X}_n - \bar{X}_n \right|^2 \right)^{\frac{1}{2}} \leq CN2^{-r}, \tag{12}$$

where $\bar{X}$ from (6) uses approximate random variables $\bar{Z} \sim \bar{\mu}_{\tilde{\mathcal{N}}}$, and C is a finite strictly positive constant.

Nested multilevel Monte Carlo I

First we perform a multilevel decomposition with different **temporal discretisations**:

$$\mathbb{E}\left(\hat{P}_L\right) = \sum_l \mathbb{E}\left(\Delta_l \hat{P}_l\right). \tag{13}$$

Then for each of these terms we form a second multilevel decomposition using the **quantised distribution**:

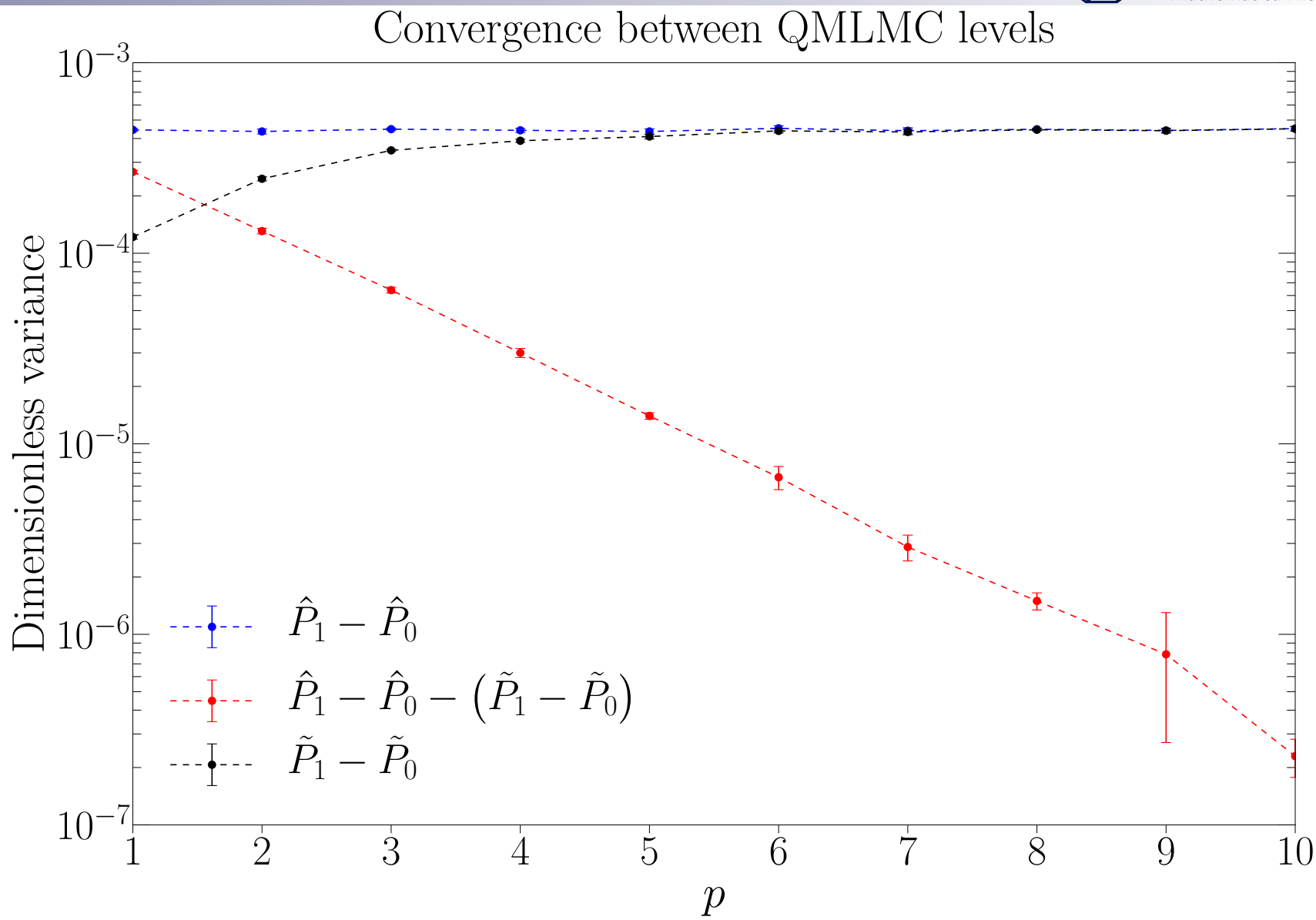
$$\mathbb{E}\left(\Delta_l \hat{P}_l\right) = \mathbb{E}\left(\Delta_l \tilde{P}_l\right) + \mathbb{E}\left(\Delta_l \hat{P}_l - \Delta_l \tilde{P}_l\right) \tag{14}$$

Nested multilevel Monte Carlo II

- Monte Carlo (MC)
  - Multilevel Monte Carlo (MLMC)
    - Quantised multilevel Monte Carlo (QMLMC)

- Monte Carlo (MC)
  - Temporal discretisation
    - Quantised distribution

Nested multilevel Monte Carlo III



Recent analysis

For the exact Euler-Scheme there are several known results which we are trying to mirror/replicate. We have shown several strong convergence results from the quantised scheme to the discrete scheme, but would ultimately prefer to compare this directly with the exact solution.

$$\mathbb{E}\left(\left|\hat{X}_T^{\mathrm{f}} - \hat{X}_T^{\mathrm{c}}\right|^p\right) \preceq \delta^{\frac{p}{2}} \tag{15}$$

$$\mathbb{E}\left(\left|\tilde{X}_T^{\mathrm{f}} - \tilde{X}_T^{\mathrm{c}}\right|^p\right) \preceq \delta^{\frac{p}{2}} \tag{16}$$

$$\mathbb{E}\left(\left|\left(\hat{X}_T^{\mathrm{f}} - \hat{X}_T^{\mathrm{c}}\right) - \left(\tilde{X}_T^{\mathrm{f}} - \tilde{X}_T^{\mathrm{c}}\right)\right|^p\right) \preceq \delta^{\frac{p}{2}} \mathbb{E}\left(\left|Z - \tilde{Z}\right|^{p(1+\epsilon)}\right)^{\frac{1}{1+\epsilon}} \tag{17}$$

$$\precapprox \delta^{\frac{p}{2}} \mathbb{E}\left(\left|Z - \tilde{Z}\right|^p\right) \tag{18}$$

Numerical results

Running our MLMC estimator with a single quantisation level (1024 bins) for finely discretised paths gives the following average time per path:

Relative accuracy $\varepsilon = 10^{-3}$ Times per path 10^{-4} s	Memory intensive	Work intensive
Original MLMC	24.8	17.0
Quantised MLMC	13.2	3.99

Level	paths
Original	1 920 000
Quantised	1 980 000
Correction	14 000

Numerical results

Running our MLMC estimator with a single quantisation level (1024 bins) for finely discretised paths gives the following average time per path:

Relative accuracy $\varepsilon = 10^{-3}$ Times per path 10^{-4} s	Memory intensive	Work intensive
Original MLMC	24.8	17.0
	$\downarrow \times 2$	
Quantised MLMC	13.2	3.99

Level	paths
Original	1 920 000
Quantised	1 980 000
Correction	14 000

Numerical results

Running our MLMC estimator with a single quantisation level (1024 bins) for finely discretised paths gives the following average time per path:

Relative accuracy $\varepsilon = 10^{-3}$ Times per path 10^{-4} s	Memory intensive	Work intensive
Original MLMC	24.8	⟶ 17.0
Quantised MLMC	13.2	⟶ 3.99

Level	paths
Original	1 920 000
Quantised	1 980 000
Correction	14 000

Numerical results

Running our MLMC estimator with a single quantisation level (1024 bins) for finely discretised paths gives the following average time per path:

Relative accuracy $\varepsilon = 10^{-3}$ Times per path 10^{-4} s	Memory intensive	Work intensive
Original MLMC	24.8	17.0
		$\downarrow \times 4$
Quantised MLMC	13.2	3.99

Level	paths
Original	1 920 000
Quantised	1 980 000
Correction	14 000

Conclusions

Errors from using a cheap proxy distribution can be **quantified** and controlled by the introduction of a nested multilevel Monte Carlo framework.

There is a degree of freedom in the construction of this proxy. Put the results in the low level **cache**, or use a very cheap (piece-wise) **polynomial**.

The resultant approximations from using such quantised distributions do **converge** under appropriate assumptions for several moments.

References I

[1] Armando Arciniega and Edward Allen. Rounding error in numerical solution of stochastic differential equations. *Stochastic Analysis and Applications*, 21(2):281–300, 2003.

[2] Steffen Omland. Mixed precision multilevel Monte Carlo algorithms for reconfigurable computing systems, June 2016.

Assumptions I

Assumption (Measurability)

The functions $a \equiv a(t,x)$ and $b \equiv b(t,x)$ are jointly (Lebesgue) $\mathcal{L}^2$-measurable in $(t,x) \in [t_0,T] \times \mathbb{R}$.

Assumption (Lipschitz continuity)

There exists a constant $K > 0$ such that $|a(t,x) - a(t,y)| \leq K|x-y|$ for all $t \in [t_0,T]$ and $x,y \in \mathbb{R}$, and identically for b.

Assumption (Bounded growth)

There exists a constant $K > 0$ such that $|a(t,x)|^2 \leq K^2(1+|x|^2)$ for all $t \in [t_0,T]$ and $x,y \in \mathbb{R}$, and identically for b.

Assumption (Measurable initial condition)

X_{t_0} is $\mathcal{F}_{t_0}$-measurable with $\mathbb{E}\left(X_{t_0}^2\right) < \infty$.

Assumptions II

Furthermore, for our Euler-Maruyama schemes with approximate solution $\hat{X}_T$ and uniform time increment δ we will further assume:

Assumption (Convergent initial value)

There exists a constant $K > 0$ such that $\mathbb{E}\left(\left|X_{t_0} - \hat{X}_{t_0}\right|^2\right) \leq K\delta$.

Assumption (Temporal Hölder continuity)

There exists a constant $K > 0$ such that $|a(t,x) - a(s,x)| \leq K(1+|x|)\sqrt{|t-s|}$ for all $s, t \in [t_0, T]$ and $x \in \mathbb{R}$, and identically for b.