Introduction to Data Assimilation
Lecture 8

Quanling Deng*

Abstract

This lecture discusses basic stochastic computational methods including the Monte Carlo method, Euler-Maruyama scheme, Milstein scheme, ensemble methods, and kernel density estimation.

Keywords Monte Carlo method, Euler-Maruyama scheme, Milstein scheme, ensemble methods, and kernel density estimation.

1 Monte Carlo Method

1.1 Basic Idea

Definition 1.1 (Monte Carlo estimator)**.** Let X be an integrable random variable with $r := \mathbb{E}[X]$. Given i.i.d. samples $X_1, \dots, X_N \sim X$, the Monte Carlo (MC) estimator is

$$\widehat{r}_N := \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Then $\widehat{r}_N$ is unbiased, $\mathbb{E}[\widehat{r}_N] = r$, and $\mathrm{Var}(\widehat{r}_N) = \mathrm{Var}(X)/N$.

Example 1.2 (MC quadrature)**.** *For* $I = \int_a^b f(x)\,dx$*, set* $U \sim \mathrm{Unif}[a,b]$*. Since* $I = (b-a)\mathbb{E}[f(U)]$,

$$\widehat{I}_N = (b-a)\frac{1}{N}\sum_{i=1}^{N} f(U_i), \qquad U_i \overset{i.i.d.}{\sim} \mathrm{Unif}[a,b].$$

For instance, computing

$$I = \int_0^1 x^2 dx = \frac{1}{3},$$

the estimates for $N = 10, 100, 1000, 10000, 100000$ are approximately

$$I = 0.20,\ 0.40,\ 0.333,\ 0.3285,\ 0.3346.$$

As N increases, $\widehat{I}_N$ converges to the true value $1/3$.
Another method is to count the ratio of random points under this curve (x, y) in the unit square.

*Yau Mathematical Science Center, Tsinghua University, Beijing, China 100048. E-mail address: qldeng@tsinghua.edu.cn

Theorem 1.3 (Error law)**.** *If* $\sigma^2 = \mathrm{Var}(X) < \infty$, *then*

$$\sqrt{N}\,(\widehat{r}_N - r) \xrightarrow{d} \mathcal{N}(0,\sigma^2), \qquad \textit{and} \quad \mathrm{SE}(\widehat{r}_N) = \sigma/\sqrt{N} = O(N^{-1/2}).$$

Let $\mu_X = \mathbb{E}[X]$ and define

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Then $\mathbb{E}[\bar{X}] = \mu_X$ and $\bar{X}$ is an unbiased estimator. The variance is

$$\mathrm{Var}(\bar{X}) = \frac{\sigma_X^2}{N},$$

where $\sigma_X^2 = \mathrm{Var}(X)$. Hence, the standard error scales as $1/\sqrt{N}$. By the central limit theorem,

$$\sqrt{N}(\bar{X} - \mu_X) \xrightarrow{d} \mathcal{N}(0,\sigma_X^2).$$

Thus, the Monte Carlo error decreases at a rate of $O(N^{-1/2})$.

Remark 1.4 (Cost of precision)**.** Halving the MC error typically requires quadrupling the sample size. To reduce the Monte Carlo uncertainty by a factor of 10, one must increase the number of samples by a factor of 100.

2 Numerical Methods for SDEs

2.1 Setting and notation

We consider the Itô SDE on $[0,T]$

$$dX_t = A(X_t,t)\,dt + B(X_t,t)\,dW_t, \qquad X_0 \in L^2, \tag{2.1}$$

where W_t is a standard m-dimensional Wiener process and $X_t \in \mathbb{R}^d$. We assume global Lipschitz and linear growth:

$$\|A(x,t) - A(y,t)\| + \|B(x,t) - B(y,t)\| \le L\|x-y\|, \tag{2.2}$$

$$x^\top A(x,t) + \tfrac{1}{2}\|B(x,t)\|_F^2 \le C(1 + \|x\|^2). \tag{2.3}$$

Let $0 = t_0 < t_1 < \cdots < t_N = T$ be a uniform grid with $\Delta t = t_{n+1} - t_n$ and $\Delta W_n := W_{t_{n+1}} - W_{t_n}$.

2.2 Generating Wiener increments in practice

Definition 2.1 (Scalar increments)**.** For scalar Brownian motion, the increments are independent Gaussians:

$$\Delta W_n \sim \mathcal{N}(0,\Delta t), \qquad \text{generate as} \quad \Delta W_n = \sqrt{\Delta t}\,Z_n,\ Z_n \sim \mathcal{N}(0,1).$$

Definition 2.2 (Multi-dimensional increments)**.** For m-dimensional W_t, take $Z_n \sim \mathcal{N}(0, I_m)$ and set $\Delta \boldsymbol{W}_n = \sqrt{\Delta t}\, Z_n$ (independent components). If a covariance $Q \in \mathbb{R}^{m\times m}$ is prescribed (correlated Wiener),

$$\Delta \boldsymbol{W}_n \sim \mathcal{N}(0, \Delta t\, Q), \qquad \Delta \boldsymbol{W}_n = \sqrt{\Delta t}\, L Z_n, \text{ with } Q = LL^\top \text{ (Cholesky).}$$

Remark 2.3 (Random seeds)**.** Fix the RNG seed when comparing schemes to ensure identical noise realizations; this isolates the *discretization* error from sampling variability.

2.3 Euler–Maruyama method and convergence

Definition 2.4 (Euler–Maruyama (EM))**.** The EM update for (2.1) is

$$X_{n+1} = X_n + A(X_n, t_n)\,\Delta t + B(X_n, t_n)\,\Delta W_n. \tag{2.4}$$

Definition 2.5 (Strong and weak errors)**.** Let $X(t_n)$ be the exact solution. The *strong error* at t_n is $e_n^{\mathrm{str}} := \big(\mathbb{E}\|X_n - X(t_n)\|^2\big)^{1/2}$. The scheme is *strong order* p if $\max_n e_n^{\mathrm{str}} = O(\Delta t^p)$. For smooth test φ,

$$e_n^{\mathrm{weak}} := \big|\mathbb{E}\,\varphi(X_n) - \mathbb{E}\,\varphi(X(t_n))\big|, \quad \text{weak order } p \iff \max_n e_n^{\mathrm{weak}} = O(\Delta t^p).$$

Theorem 2.6 (Moment bounds)**.** *Under* (2.2)–(2.3)*,* $\sup_{t\in[0,T]} \mathbb{E}\|X(t)\|^2 < \infty$ *and for EM,* $\max_n \mathbb{E}\|X_n\|^2 \le C$ *uniformly in* Δt *(for small enough* Δt*).*

Proof sketch. Apply Itô's formula to $\|X_t\|^2$ and use (2.3), then Grönwall. For EM, expand $\|X_{n+1}\|^2$ via (2.4), take expectations, use $\mathbb{E}[\Delta W_n] = 0$, $\mathbb{E}\|\Delta W_n\|^2 = m\Delta t$, and discrete Grönwall. □

Theorem 2.7 (Strong order of EM)**.** *Under* (2.2)–(2.3)*, EM has strong order* $1/2$*:*

$$\max_{0\le n\le N} \big(\mathbb{E}\|X_n - X(t_n)\|^2\big)^{1/2} \le C\,\Delta t^{1/2}.$$

Details. Write the exact variation-of-constants over $[t_n, t_{n+1}]$:

$$X(t_{n+1}) = X(t_n) + \underbrace{\int_{t_n}^{t_{n+1}} A(X_s, s)\,ds}_{D_n} + \underbrace{\int_{t_n}^{t_{n+1}} B(X_s, s)\,dW_s}_{M_n}.$$

Subtract EM (2.4) and define $E_n := X(t_n) - X_n$:

$$E_{n+1} = E_n + \big(D_n - A(X_n, t_n)\Delta t\big) + \big(M_n - B(X_n, t_n)\Delta W_n\big).$$

Decompose $D_n = \int_{t_n}^{t_{n+1}} [A(X_s, s) - A(X_n, t_n)]\,ds + A(X_n, t_n)\Delta t$, so the $A(X_n, t_n)\Delta t$ cancels, and similarly for M_n. Hence

$$E_{n+1} = E_n + R_n^D + R_n^M,$$

with

$$R_n^D := \int_{t_n}^{t_{n+1}} \big(A(X_s,s) - A(X_n,t_n)\big)ds, \qquad R_n^M := \int_{t_n}^{t_{n+1}} \big(B(X_s,s) - B(X_n,t_n)\big)dW_s.$$

Take squared norms and expectations, use $(a+b+c)^2 \le (1+\eta)a^2 + C_\eta(b^2+c^2)$ for any $\eta > 0$:

$$\mathbb{E}\|E_{n+1}\|^2 \le (1+\eta)\,\mathbb{E}\|E_n\|^2 + C_\eta\big(\mathbb{E}\|R_n^D\|^2 + \mathbb{E}\|R_n^M\|^2\big).$$

By Lipschitz and Jensen,

$$\|A(X_s,s) - A(X_n,t_n)\| \le L\|X_s - X_n\| + L|s-t_n|.$$

Using moment bounds (Thm. 2.6) and standard Hölder continuity of X_t in L^2 ($\mathbb{E}\|X_s - X_{t_n}\|^2 \le C|s-t_n|$), one gets

$$\mathbb{E}\|R_n^D\|^2 \le C\,(\Delta t)^2.$$

For the martingale remainder, BDG inequality with Lipschitz gives

$$\mathbb{E}\|R_n^M\|^2 \le C\,\mathbb{E}\int_{t_n}^{t_{n+1}} \|B(X_s,s) - B(X_n,t_n)\|^2 ds \le C\,(\Delta t)^2 + C\,\Delta t\,\mathbb{E}\|E_n\|^2.$$

Combine:

$$\mathbb{E}\|E_{n+1}\|^2 \le (1 + C\Delta t)\,\mathbb{E}\|E_n\|^2 + C\,(\Delta t)^2.$$

Discrete Grönwall yields $\mathbb{E}\|E_n\|^2 \le C\,\Delta t$, i.e. strong order 1/2. □

Theorem 2.8 (Weak order of EM)**.** *If A, B and test φ are sufficiently smooth with polynomial growth, then EM has weak order* 1*:*

$$\max_n \big|\mathbb{E}\,\varphi(X_n) - \mathbb{E}\,\varphi(X(t_n))\big| \le C\,\Delta t.$$

Idea (Talay–Tubaro expansion). Let $u(x,t) = \mathbb{E}\big[\varphi(X_T) \mid X_t = x\big]$, so u solves the backward Kolmogorov PDE

$$\partial_t u + \mathcal{L}u = 0, \qquad \mathcal{L}\psi = A\cdot\nabla\psi + \tfrac{1}{2}(BB^\top) : \nabla^2\psi.$$

Compute one-step weak error

$$\epsilon_n := \big|\mathbb{E}\,u(X_{n+1}, t_{n+1}) - \mathbb{E}\,u(X(t_{n+1}), t_{n+1})\big|.$$

A second-order Itô–Taylor expansion of $u(X_{n+1}, t_{n+1})$ around (X_n, t_n) and matching with the generator gives $\epsilon_n \le C\,(\Delta t)^2$ uniformly. Summation over $N = T/\Delta t$ steps yields $O(\Delta t)$. □

2.4 Milstein method and convergence

Definition 2.9 (Milstein (scalar noise, $m = 1$))**.** For sufficiently smooth $A, B : \mathbb{R}^d \to \mathbb{R}^d$, the Milstein update is

$$X_{n+1} = X_n + A(X_n)\Delta t + B(X_n)\Delta W_n + \tfrac{1}{2}\big(\nabla B(X_n)\, B(X_n)\big)\big((\Delta W_n)^2 - \Delta t\big), \qquad (2.5)$$

where $(\nabla B)B$ denotes the directional derivative (Jacobian of B applied to B). For multi-dimensional noise, Lévy area terms appear (omitted here for brevity).

Theorem 2.10 (Strong order of Milstein)**.** *Under global Lipschitz and C^2 smoothness of A, B, the Milstein method* (2.5) *has strong order* 1*:*

$$\max_n \big(\mathbb{E}\|X_n - X(t_n)\|^2\big)^{1/2} \le C\,\Delta t.$$

Details. Apply the Itô–Taylor expansion up to terms of order Δt:

$$X(t_{n+1}) = X(t_n) + A(X_n)\Delta t + B(X_n)\Delta W_n + \tfrac{1}{2}(\nabla B\,B)(X_n)\big((\Delta W_n)^2 - \Delta t\big) + R_n,$$

with $\mathbb{E}\|R_n\|^2 \le C\,(\Delta t)^3$. Subtract (2.5), set $E_n = X(t_n) - X_n$, and proceed as in Theorem 2.7. The leading local mean-square error is $O(\Delta t^{3/2})$, giving global strong $O(\Delta t)$. □

Theorem 2.11 (Weak order of Milstein)**.** *Under smoothness assumptions as above, Milstein is weak order* 1 *(same as EM in the scalar-noise case).*

Remark 2.12 (When Milstein helps)**.** Milstein improves *strong* order from 1/2 (EM) to 1 (scalar noise) without changing the weak order. This is beneficial when path-wise accuracy matters (e.g. strong DA constraints, path-dependent payoffs). For multi-dimensional noise, Milstein requires Lévy areas or approximations (Kloeden–Platen), which is an excellent talking point about complexity vs. accuracy.

Remark 2.13 (Constant diffusion)**.** If B is constant, the Milstein correction vanishes and Milstein reduces to EM; in that case strong order is already 1 (since diffusion is additive).

2.5 Mathematical Tools

Theorem 2.14 (Burkholder–Davis–Gundy (BDG) Inequality)**.** *Let $(M_t)_{t\ge 0}$ be a continuous local martingale with $M_0 = 0$ and quadratic variation process*

$$\langle M\rangle_t = \int_0^t |H_s|^2\,ds,$$

where H_s is a predictable (adapted) process. Then for any $p \ge 1$, there exist constants C_p, $C_p' > 0$ such that

$$C_p^{-1}\,\mathbb{E}\Big[\langle M\rangle_T^{p/2}\Big] \le \mathbb{E}\Bigg[\sup_{0\le t\le T}|M_t|^p\Bigg] \le C_p'\,\mathbb{E}\Big[\langle M\rangle_T^{p/2}\Big]. \qquad (2.6)$$

In particular, for stochastic integrals of the form

$$M_t = \int_0^t H_s \, dW_s,$$

we have

$$\mathbb{E}\left[\sup_{0 \le t \le T} \left| \int_0^t H_s \, dW_s \right|^p\right] \le C_p \, \mathbb{E}\left[\left(\int_0^T |H_s|^2 \, ds\right)^{p/2}\right].$$

Proof idea. The result is nontrivial and relies on martingale inequalities and stopping-time arguments. The intuition is that the L^p-norm of the maximal fluctuation of a martingale M_t is equivalent (up to constants) to the L^p-norm of its quadratic variation $\langle M \rangle_T^{1/2}$, i.e. the total accumulated variance. □

Remark 2.15 (Practical use in SDE analysis)**.** In SDE convergence proofs, the BDG inequality is typically applied to stochastic remainder terms such as

$$R_n^M = \int_{t_n}^{t_{n+1}} \big(B(X_s, s) - B(X_n, t_n)\big) \, dW_s.$$

Taking expectations and applying BDG with $p = 2$ gives

$$\mathbb{E}\|R_n^M\|^2 \le C \, \mathbb{E}\left[\int_{t_n}^{t_{n+1}} \|B(X_s, s) - B(X_n, t_n)\|^2 ds\right].$$

This converts the difficult stochastic integral into a deterministic time integral, which can then be bounded using Lipschitz continuity and moment estimates of X_t.

Remark 2.16 (Historical note)**.** The inequality is named after Donald Burkholder, Benjamin Davis, and Ronald Gundy, who established it in the 1960s. It's a beautiful generalization of Doob's maximal inequality, connecting martingale oscillations with their "energy." In stochastic numerics, BDG is one of the main tools to control $\mathbb{E}\|X_t - X_n\|^p$.

Theorem 2.17 (Discrete Grönwall Inequality)**.** *Let $\{a_n\}, \{b_n\}, \{c_n\}$ be nonnegative sequences satisfying*

$$a_n \le b_n + \sum_{k=0}^{n-1} c_k \, a_k, \qquad n \ge 1. \tag{2.7}$$

If $c_k \le C$ for all k, then

$$a_n \le b_n + C \sum_{k=0}^{n-1} b_k \, e^{C(n-k-1)}. \tag{2.8}$$

In particular, if $b_n = C_1 \Delta t$ and $c_n = C_2 \Delta t$, then

$$a_n \le C_1 \Delta t \sum_{k=0}^{n-1} (1 + C_2 \Delta t)^{n-k-1} \le \frac{C_1}{C_2}\big(e^{C_2 n \Delta t} - 1\big),$$

so $a_n = O(\Delta t)$ uniformly in n for bounded $T = n\Delta t$.

Proof idea. The proof mirrors the continuous Grönwall inequality. Starting from (2.7), define $d_n = \sum_{k=0}^{n-1} c_k$. Then by induction:

$$a_n \le b_n + \sum_{k=0}^{n-1} b_k \prod_{j=k+1}^{n-1} (1 + c_j).$$

Using the elementary bound $(1 + c_j) \le e^{c_j}$ gives

$$a_n \le b_n + \sum_{k=0}^{n-1} b_k e^{\sum_{j=k+1}^{n-1} c_j} \le b_n + e^{d_n} \sum_{k=0}^{n-1} b_k,$$

and if $c_j \le C$, then $e^{d_n} \le e^{Cn}$, yielding (2.8). □

Remark 2.18 (Intuitive meaning)**.** The inequality bounds any recursively growing quantity a_n that depends linearly on its previous values and a known source term b_n. It prevents exponential blow-up provided c_n is controlled. In convergence proofs (e.g. Euler–Maruyama), it ensures that local errors of size $O(\Delta t^p)$ accumulate only linearly, leading to global errors of order $O(\Delta t^{p-1})$ or $O(\Delta t^p)$, depending on context.

3 Ensemble Methods and Kernel Density Estimation

3.1 Ensemble simulation

Definition 3.1 (Ensemble Monte Carlo for SDEs)**.** Run M independent copies using the chosen scheme (EM or Milstein) and independent Wiener increments:

$$X_{n+1}^{(i)} = X_n^{(i)} + A(X_n^{(i)}, t_n)\Delta t + B(X_n^{(i)}, t_n)\Delta W_n^{(i)}, \qquad i = 1, \dots, M.$$

The ensemble mean and covariance approximate $\mathbb{E}[X(t_n)]$ and $\mathrm{Cov}[X(t_n)]$:

$$\widehat{\mu}_n = \frac{1}{M} \sum_{i=1}^{M} X_n^{(i)}, \qquad \widehat{R}_n = \frac{1}{M-1} \sum_{i=1}^{M} \big(X_n^{(i)} - \widehat{\mu}_n\big)\big(X_n^{(i)} - \widehat{\mu}_n\big)^{\top}.$$

Remark 3.2 (Ensemble size vs. time step)**.** Emphasize the trade-off: for a fixed compute budget, there is an optimal balance between decreasing Δt (discretization error) and increasing M (sampling error).

Remark 3.3 (Ergodic systems and ensemble analysis)**.** In many stochastic dynamical systems, such as Langevin equations or dissipative stochastic oscillators, the process $\{X_t\}_{t \ge 0}$ is *ergodic.* This means that there exists a unique invariant (stationary) probability distribution $\pi(x)$ such that, for any suitable observable $\varphi(x)$,

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \varphi(X_t)\, dt = \int_{\mathbb{R}^d} \varphi(x)\, \pi(x)\, dx \quad \text{almost surely.}$$

Consequently, one can approximate ensemble expectations by long-time averages along a single trajectory:

$$\mathbb{E}_\pi[\varphi(X)] \approx \frac{1}{T}\int_0^T \varphi(X_t)\,dt,$$

which is often referred to as the *ergodic theorem*.

In practice, this implies that for ergodic systems, we can replace the large ensemble $\{X_t^{(i)}\}_{i=1}^M$ with a single sufficiently long trajectory, significantly reducing computational cost. However, ergodicity also requires adequate *mixing*—the trajectory must explore the entire state space representative of $\pi(x)$ within the simulation horizon. Poor mixing (e.g. in systems with multiple metastable states) leads to biased estimates, since the trajectory may remain trapped in one region of the phase space.

- **Ergodic average:** time average equals ensemble average.
- **Mixing rate:** determines how fast the process "forgets" its initial condition and converges to equilibrium.
- **Practical implication:** in ensemble-based data assimilation, ergodicity ensures that sampling across long runs is statistically equivalent to sampling across multiple realizations.

3.2 Kernel density estimation (KDE)

Kernel density estimation (KDE) provides a smooth, non-parametric way to approximate the probability density function (PDF) from a finite ensemble. Given samples $\{x^{(i)}\}_{i=1}^N$ from a univariate distribution with density p, the kernel density estimator is

$$\widehat{p}_h(x) = \frac{1}{Nh}\sum_{i=1}^N K\left(\frac{x - x^{(i)}}{h}\right),$$

with kernel K (e.g. Gaussian $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$) and bandwidth $h > 0$. This can be viewed as a local averaging (or smoothing) of Dirac delta functions centered at the ensemble points. For Gaussian-like data with sample standard deviation $\hat{\sigma}$,

$$h_{\mathrm{rot}} \approx 1.06\,\hat{\sigma}\,M^{-1/5}.$$

Remark 3.4 (Multivariate KDE)**.** For d-dimensional X, use

$$\widehat{p}_H(\boldsymbol{x}) = \frac{1}{M}\sum_{i=1}^M \frac{1}{\sqrt{(2\pi)^d \det H}} \exp\Big(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(i)})^\top H^{-1}(\boldsymbol{x} - \boldsymbol{x}^{(i)})\Big),$$

with a positive-definite bandwidth matrix H. In DA, H is often chosen proportional to the ensemble covariance.