Part III: discrete-time dynamics of diffusions and discrete-time likelihood

- Transition density, discrete-time likelihood, approximate dynamics based on discretizations
- Pseudo-likelihood approaches
- Exact simulation of diffusions using rejection sampling on the path space (Girsanov and a transformation)

Transition density and discrete-time likelihood

We have seen the Markov transition operator which determines the **exact macroscopic dynamics**; its Lebesgue density is called the transition density:

$$p_{s,t}(v, w; \theta) = \mathbb{P}\left[V_t \in \mathrm{d}w \mid V_s = v \right] / \mathrm{d}w, \quad t > s,\ w, v \in R^d . \tag{23}$$

For time homogenous diffusions, this only depends on $t - s$, and we simplify the notation

In statistical terms, the transition density gives the density of the **conditional distribution** of V_t given $V_s = v$.

Discrete-time likelihood

Therefore, the joint density of a *discretely observed diffusion*, $\{V_{t_0}, V_{t_1}, \ldots, V_{t_n}\}$, due to the Markov property, is simply given by the product of the transition densities

$$L(\theta \mid \mathbf{v}) = \prod_{i=0}^{n-1} p_{t_i, t_{i+1}}(V_{t_i}, V_{t_{i+1}}; \theta)\,. \tag{24}$$

However, the transition density is rarely available in closed form. Exceptions are some of the linear SDEs we have seen (e.g 1-d with additive error, multi-d with additive error and constant coefficients), or the so-called CIR model [Cox et al., 1985]

We will see later insights as to why it is intractable. Mathematically, however, it has a very clean representation

Focusing on time-homogeneous diffusions, $p_t(v, w)$ with w fixed, as a function of (v, t) solves the Kolmogorov backward equation

$$\frac{\partial p}{\partial t} = Ap \tag{25}$$

with initial condition $p_T(v, w)$ is a Dirac-delta centred at w, whereas with v fixed , as a function of (w, t) solves the Kolmogorov forward equation

$$\frac{\partial p}{\partial t} = A^* p \tag{26}$$

with initial condition $p_0(v, w)$ is a Dirac-delta centred at v, where A is the diffusion **generator** and A^* its adjoint operator, e.g for 1-d diffusions this is given by

$$A^* f = -(bf)' + \frac{1}{2}(\sigma^2 f)''$$

(there might be additional boundary conditions)

In the context of the discrete-time dynamics, it is the KFE that we are interested in.

However, the KFE can be solved exactly in the cases mentioned earlier, which are very limited. Numerical solutions based on standard PDE techniques is a possibility; see the recent article [Hurn et al., 2007] and the tutorial [Aït-Sahalia et al., 2004] for operator methods in this context.

Note that such approaches require space-time discretizations. Additionally, for statistical purposes we will need several such evaluations for different t's and different pairs (v, w).

Therefore, Monte Carlo methods offer a very attractive alternative for approximating the discrete-time dynamics of diffusions.

Approximate simulation of diffusion skeletons

Traditional methods involve time-discretisation of the SDE in order to obtain an approximation to the discrete-time dynamics of the diffusion. See [Kloeden and Platen, 1995] for an extensive treatment.

The simplest and quickest is the Euler-Maruyama method:

$$V_{t+\Delta} = V_t + \Delta b(V_t) + \sigma(V_t) Z_\Delta \tag{27}$$

where Z_Δ is a zero-mean, unit-variance random variable. This gives the first clue why transition density is intractable: non-linear convolution of Gaussians.
How might we assess the approximation of this and other methods? Strong approximation of order γ

$$\mathbb{E}[|V_T - V_T^\Delta|] \leq K\Delta^\gamma$$

Weak approximation of order γ

$$|\mathbb{E}[g(V_T)] - \mathbb{E}[g(V_T^\Delta)]| \leq K\Delta^\gamma$$

for suitable functions g.

A strong Euler method uses $Z_\Delta = B_{t+\Delta} - B_t$. Under suitable regularity, strong Euler is of order $1/2$ in the strong sense, whereas in general Euler schemes are of order 1 in the weak sense.

Many higher order schemes exist. Many are based on the Itô-Taylor expansion. Recall that

$$f(V_t) = f(V_0) + \int_0^t A^0 f(V_s)\mathrm{d}s + \int_0^t A^1 f(V_s)\mathrm{d}B_s$$

where $A^0 = A$, $A^1 = \sigma(x)\mathrm{d}/\mathrm{d}x$.

We can perform a further Itô expansion of both $A^0 f$ and $A^1 f$

$$f(V_t) = f(V_0) + A^0 f(V_0)t + A^1 f(V_0)B_t + R$$

where

$$R = \int_0^t \int_0^s A^{00} f(V_r) \mathrm{d}r\mathrm{d}s + \int_0^t \int_0^s A^{10} f(V_r) \mathrm{d}B_r\mathrm{d}s$$

$$+ \int_0^t \int_0^s A^{01} f(V_r) \mathrm{d}r\mathrm{d}B_s + \int_0^t \int_0^s A^{11} f(V_r) \mathrm{d}B_r\mathrm{d}B_s$$

and $A^{ij} f = A^i(A^j f)$.
Applying this for $f(v) = v$ we get the strong Euler scheme, and R gives an explicit form for the error

R contains four terms. For small time intervals, Brownian fluctuations dominate drift terms ($\mathrm{d}B_r \sim O(\mathrm{d}r^{1/2})$) so that the fourth term dominates:

$$\int_0^t \int_0^s A^{11} f(V_r) \mathrm{d}B_r \mathrm{d}B_s$$

For f the identity function the fourth error term reduces to

$$\int_0^t \int_0^s \sigma(V_r)\sigma'(V_r) \mathrm{d}B_r \mathrm{d}B_s$$

Assuming that σ is continuously differentiable, then for small r, $\sigma(V_r)\sigma'(V_r) \approx \sigma(V_0)\sigma'(V_0)$, and error term is

$$\sigma(V_0)\sigma'(V_0) \int_0^t \int_0^s \mathrm{d}B_r \mathrm{d}B_s = \sigma(V_0)\sigma'(V_0) \frac{(B_t^2 - t)}{2} \ .$$

This leads to the Milstein approximation scheme

$$V_{t+\Delta} = V_t + b(V_t)\Delta + \sigma(V_t)Z_\Delta + \frac{\sigma(V_t)\sigma'(V_t)}{2}(Z_\Delta^2 - \Delta) \ .$$

The Milstein scheme is a strong order 1 approximation of the diffusion.
Higher order Itô-Taylor expansions further expand some or all of the four terms in R by Itô's formula.

Other ways of improving on Euler-Maruyama exist, including implicit and split-step methods (which are particularly important in the construction of MCMC methods using diffusion dynamics).

The following family of approximation is particularly relevant in statistics for diffusions; it directly has a flavour of the Extended Kalman Filter, but it turns out that it is instrumental in MC methods for diffusions. Additionally, an empirical investigation in [Durham and Gallant, 2002] shows that it is among the most accurate of the approximation methods

We present it for univariate diffusions. The main idea is that since we can solve explicitly linear SDEs with additive error, we can **locally** approximate a non-linear SDE with a linear and solve it.

Local linearization

The idea is at a time increment $[t, t+\Delta]$ approximate the drift in the SDE for V by a linear function and the diffusion by a constant, and solve explicitly the SDE to obtain $V_{t+\Delta}$ as a **locally** Gaussian random variable whose mean depends on V_t. There are various ways to attempt the linearization.

Note that the Euler scheme is precisely a form of local linearization where the both coefficients are replaced by constants

Central to linear approximation is Taylor expansion...

Using Itô on $b(t, V_t)$ to get:

$$\mathrm{d}b(s, V_s) = \frac{\partial b}{\partial s}(s, V_s) + \frac{\partial b}{\partial v}(s, V_s)\mathrm{d}V_s + \frac{1}{2}\frac{\partial^2 b}{\partial v^2}\sigma^2(s, V_s)\mathrm{d}s$$

Assuming that $\frac{\partial^2 b}{\partial v^2}\sigma^2(s, V_s)$, $\frac{\partial b}{\partial s}$ and $\frac{\partial b}{\partial v}$ are constant in $[t, t + \Delta]$ we obtain that for s in the interval,

$$\begin{aligned}
\tilde{b}(s, V_s) &= \left(\frac{\partial b}{\partial t}(t, V_t) + \frac{1}{2}\frac{\partial^2 b}{\partial v^2}\sigma^2(t, V_t)\right) s + \frac{\partial b}{\partial v}(t, V_t)V_s \\
&+ b(t, V_t) - \frac{\partial b}{\partial v}(t, V_t)V_t - \left(\frac{\partial b}{\partial t}(t, V_t) - \frac{1}{2}\frac{\partial^2 b}{\partial v^2}\sigma^2(t, V_t)\right) t \\
&= L_t V_s + M_t s + N_t
\end{aligned}$$

and $\tilde{\sigma}(s, V_s) = \sigma(t, V_t)$

Thus we obtain an approximating SDE

$$\mathrm{d}V_s = (L_t V_s + M_t s + N_t)\mathrm{d}s + \tilde{\sigma}_t \mathrm{d}B_s \,, s \in [t, t+\Delta], \qquad (28)$$

which can be solved analytically to yield $V_{t+\Delta}$ as a (non-linear) function of V_t, yielding a **locally conditional Guassian** approximating transition density

The approach for constant σ has been developed in a sequence of articles by Ozaki, and Shoji and Ozaki, see for example [Ozaki, 1992], [Shoji and Ozaki, 1998], and the weak-strong error has been investigated. Nevertheless, the more general localization described is a valid possibility for general σ.

Statistical use of discretizations: pseudo-likelihood

The approximations to the discrete-time dynamics can be used within pseudo-likelihood approach.

As with the exact solutions, we are interested in discretizations which yield explicit conditional density for consecutive values. Such examples are the Euler and the local linearization. The resulting approximation is applied to each pair of observations v_i, v_{i+1}, to yield the approximation $\hat{p}_{t_i,t_{i+1}}(v_i, v_{i+1}; \theta)$, and the approximation to (24)

$$\hat{L}(\theta \mid \mathbf{v}) = \prod_{i=0}^{n-1} \hat{p}_{t_i,t_{i+1}}(v_i, v_{i+1}; \theta)\,.$$

Such approach has been undertaken e.g in [Ozaki, 1992], using the localization approach, but a common unpleasant feature of these pseudo-likelihood approaches is inconsistency; this has been proved in [Florens-Zmirou, 1989]. The asymptotic considered is keeping distance fixed, and increasing number of observations (outfill asymptotics)

As a simple example, you can consider the Ornstein-Uhlenbeck process, and compare the MLE for σ and ϕ with the pseudo-MLE; the former are consistent as $n \to \infty$ but the latter not, see for example [Pedersen, 1995]

The same is true of other pseudo-likelihood approaches, e.g assume that $V_{t+\Delta}$ conditionally on V_t is Gaussian with an approximation to the mean and variance; see for example [Kessler, 1997]. A direct approximation is given by the Euler, in which case the approach collapses to the approach described earlier.

However, better approximations can be obtained, e.g based on the ODEs solved by moments of the diffusion, recall the KFE. The inconsistency is a result of the mispecification of the moments, lack of efficiency due to the assumed Gaussianity. Note however, that under certain conditions, a correct specification of the moments would yield consistent estimators.

The requirement to obtain statistically consistent and efficient estimation procedures has motivated the development of MC methods in this framework. Before proceeding, however, lets see a different exact representation of the discrete-time dynamics for SDEs for which the solution is intractable.

Exact simulation of diffusions

We will present the idea in the simplest framework for 1-d, time-homegenous diffusions. Various extensions exist. The approach has been developed in detail in [Beskos et al., 2006a, Beskos et al., 2008, Beskos et al., 2006b].

This is another use of Girsanov's theorem and the likelihood ratio on the path space in inference for diffusions. Recall the canonical representation for the Brownian motion.

An important tool: Brownian bridge

An indispensable component of exact simulation methods, but more generally of MC for diffusions (see later) is the stochastic process known as the Brownian bridge. It solves the linear SDE

$$\mathrm{d}X_s = \frac{y - X_s}{T - s}\mathrm{d}s + \mathrm{d}B_s\,, s \in [0, T] \tag{29}$$

and has macroscopic dynamics specified, for $0 < t_1 < t_2 < T$, as

$$X_{t_2} \mid X_{t_1} \sim N\left(X_{t_1} + \frac{t_2 - t_1}{T - t_1}(y - V_{t_1}),\ \frac{(t_2 - t_1)(T - t_2)}{T - t_1} I_d\right). \tag{30}$$

It can be shown from first principles, that this process has the same law as that of Brownian motion **conditioned** at its end point $B_T = y$. It is the first instance of a general class of processes we will see later, known as diffusion bridges

Rejection sampling

Recall the basic principle of rejection sampling, presented here in a **very generic way**

Probability space of interest $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{P}$ the target measure, difficult to simulate from.
$\mathbb{W}$ another measure on $(\Omega, \mathcal{F})$ easy to simulate from, s.t the **likelihood ratio** (Radon-Nikodym derivative) exists and

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{W}}(\omega) := f(\omega) \leq K < \infty\,, \quad \forall \omega \in \Omega$$

Then, the algorithm for generating one draw proceeds as follows

1. Propose $\omega \sim \mathbb{W}$
2. **Generate a coin whose probability of heads is** $f(\omega)/K$
3. If heads, accept and store ω, otherwise return to 1.

Note the freedom in 2, it is not really necessary to carry out in the "traditional way" of $U \sim Uni[0,1]$ and comparing it to $f(\omega)/K$.

Let $\mathbb{P}^{(t,v)}$ denote the probability measure generated by the solution of

$$\mathrm{d}V_s = b(V_s)\mathrm{d}s + \sigma(V_s)\mathrm{d}B_s\,, \quad s \in [0,t], V_0 = v \tag{31}$$

Then, $V_t \sim p_t(v, \mathrm{d}w)$ is a draw from **the marginal** of $\mathbb{P}^{(t,v)}$

Therefore, instead of simulating V_t we can think of the - at first view - much harder task of simulating the whole diffusion path

Nevertheless, we should look for a measure on the path space, from which we can at least simulate paths (in a sense to be made precise, for the moment lets think of high-frequency skeletons)

But, from the quadratic variation identity we know that such measure will have the same diffusion coefficient as (31), and from the discussion on solvable SDEs we know that unless σ is constant it would be in general impossible to obtain samples from any such process. Nevertheless, Itô can help us to change (31) into a process of **unit diffusion**, for which we have a much better chance to simulate exactly

Lamperti transformation

$$\mathrm{d}V_s = b(V_s)\mathrm{d}s + \sigma(V_s)\mathrm{d}B_s$$

Transform $V_s \to \eta(V_s) =: X_s$ where

$$\eta(u) = \int^u \frac{1}{\sigma(z)}\,\mathrm{d}z \tag{32}$$

Itô's rule: $\mathrm{d}X_s = \alpha(X_s)\,\mathrm{d}s + \mathrm{d}B_s, \quad X_0 = x := \eta(V_0),\ s \in [0, t],$

where

$$\alpha(u;\theta) = \frac{b\{\eta^{-1}(u;\theta);\theta\}}{\sigma\{\eta^{-1}(u;\theta);\theta\}} - \sigma'\{\eta^{-1}(u;\theta);\theta\}\,/\,2,$$

Exact simulation of V is equivalent to exact simulation of X

$$\mathrm{d}X_s = \alpha(X_s)\,\mathrm{d}s + \mathrm{d}B_s, \quad s \in [0,t], X_0 = x \tag{33}$$

Let $\mathbb{Q}^{(t,x)}$ be the distribution on path space induced by X, and $\mathbb{W}^{(t,x)}$ the correspond Wiener measure.

By Girsanov's formula:

$$\frac{\mathrm{d}\mathbb{Q}^{(t,x)}}{\mathrm{d}\mathbb{W}^{(t,x)}}(\omega) = \exp\left\{\int_0^t \alpha(\omega_s)\mathrm{d}\omega_s - \frac{1}{2}\int_0^t \alpha^2(\omega_s)\mathrm{d}s\right\}$$

Another use of Itô: elimination of stochastic integrals Let,

$$A(u) := \int^u \alpha(z)\mathrm{d}z,$$

$$\frac{\mathrm{d}\mathbb{Q}^{(t,x)}}{\mathrm{d}\mathbb{W}^{(t,x)}}(\omega) = \exp\left\{A(\omega_t) - A(x) - \frac{1}{2}\int_0^t (\alpha^2 + \alpha')(\omega_s)\mathrm{d}s\right\}$$

Pause for thought

Aim at rejection sampling from $\mathbb{Q}^{(t,x)}$ using proposals from (something close) to $\mathbb{W}^{(t,x)}$

Girsanov's formula gives the density ratio (which is also a likelihood ratio)

Density ratio needs to be bounded, and we have still to figure out how to avoid infinite simulation

Biased Brownian proposals

Process with ending point density (assumed to be integrable)

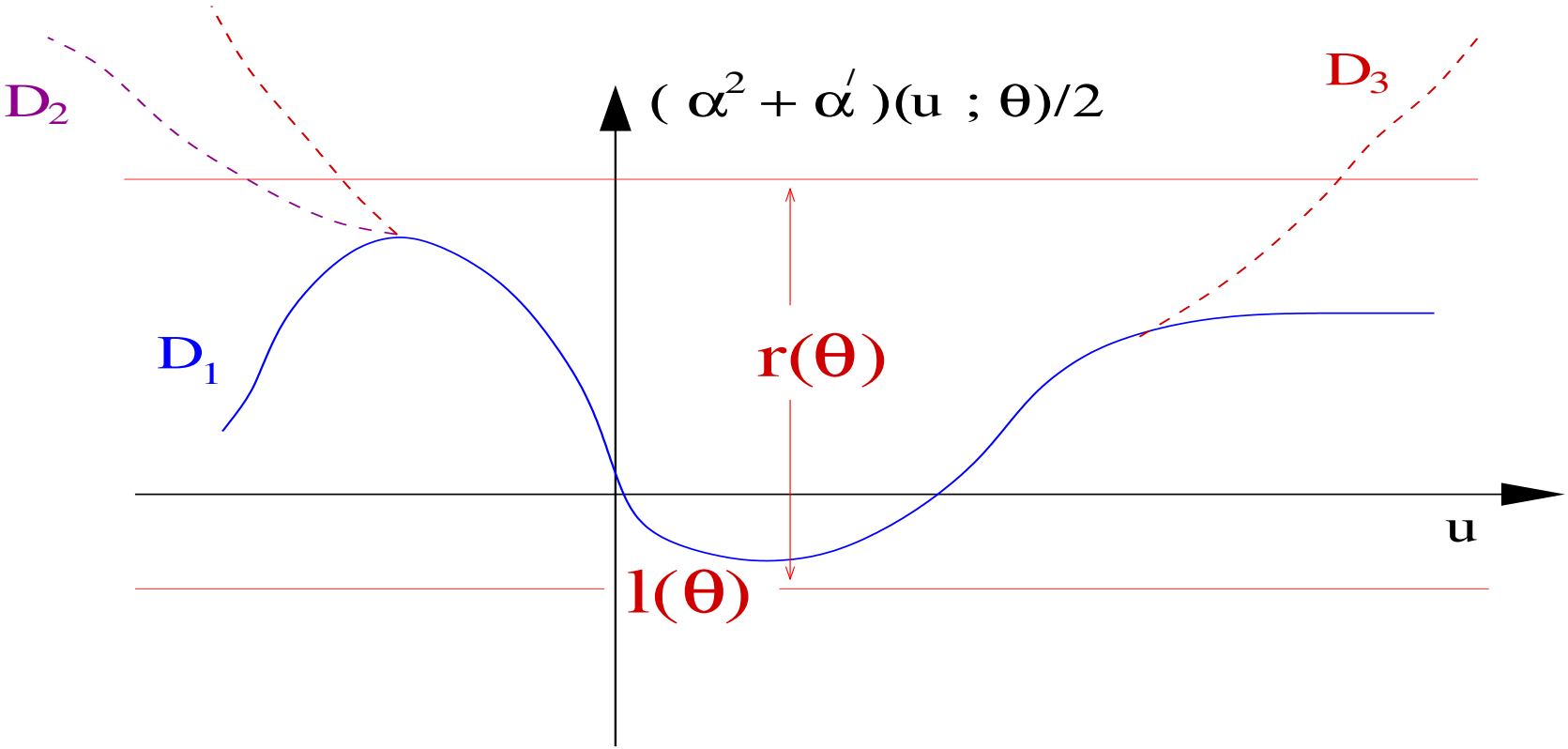
$$h(u) \propto \exp\{A(u) - (u-x)^2/2t\}, \quad u \in \mathbf{R}$$

Thus, $\omega_t \sim h$, and $(\omega_s, 0 \leq s < t) \mid \omega_t$, from Brownian bridge.
$\mathbb{Z}^{(t,x)}$ the distribution induced by the biased Brownian motion.
Then

$$\frac{\mathrm{d}\mathbb{Q}^{(t,x)}}{\mathrm{d}\mathbb{Z}^{(t,x)}}(\omega) \propto \exp\left\{-\int_0^t \frac{1}{2}(\alpha^2 + \alpha')(\omega_s)\mathrm{d}s\right\}$$

Summary of assumptions

1. The drift function α is differentiable
2. The function $\exp\{A(u) - (u-x)^2/2t\}$, $u \in \mathbf{R}$, is integrable
3. The function $(\alpha^2 + \alpha')(\cdot)$ is bounded (this is working assumption for the lecture, it can be relaxed)



Example: Periodic drift $\alpha(u;\theta) = \sin(u-\theta)$

$$
\begin{aligned}
l &\leq \inf_{u \in \mathbf{R}} \{(\alpha^2 + \alpha')(u)/2\}, \quad \textit{minimum} \\
r &\geq \sup_{u \in \mathbf{R}} \{(\alpha^2 + \alpha')(u)/2 - l\}, \quad \textit{range}
\end{aligned}
$$

Define non-negative $0 \leq \phi \leq 1$:

$$
\phi(\omega_s) = \frac{1}{r} \left\{ \frac{(\alpha^2 + \alpha')(\omega_s)}{2} - l \right\}, \quad s \in [0, t]
$$

Bounded likelihood ratio

$$\frac{\mathrm{d}\mathbb{Q}^{(t,x)}}{\mathrm{d}\mathbb{Z}^{(t,x)}}(\omega) \propto \exp\left\{-r\int_0^t \phi(\omega_s)\mathrm{d}s\right\} \leq 1, \quad \mathbb{Z}^{(t,x)} - \mathrm{a.s.}.$$

Thought process so far

- Wish to simulate V_t given $V_0 = v$
- Embed this into simulating $(V_s, s \in [0, t])$ given $V_0 = v$. Dead-end
- Transform $V \to X$, simulate X-path using rejection sampling
- Find a dominating measure from which it is easy to simulate and which has bounded LR wrt to that of X. We have it

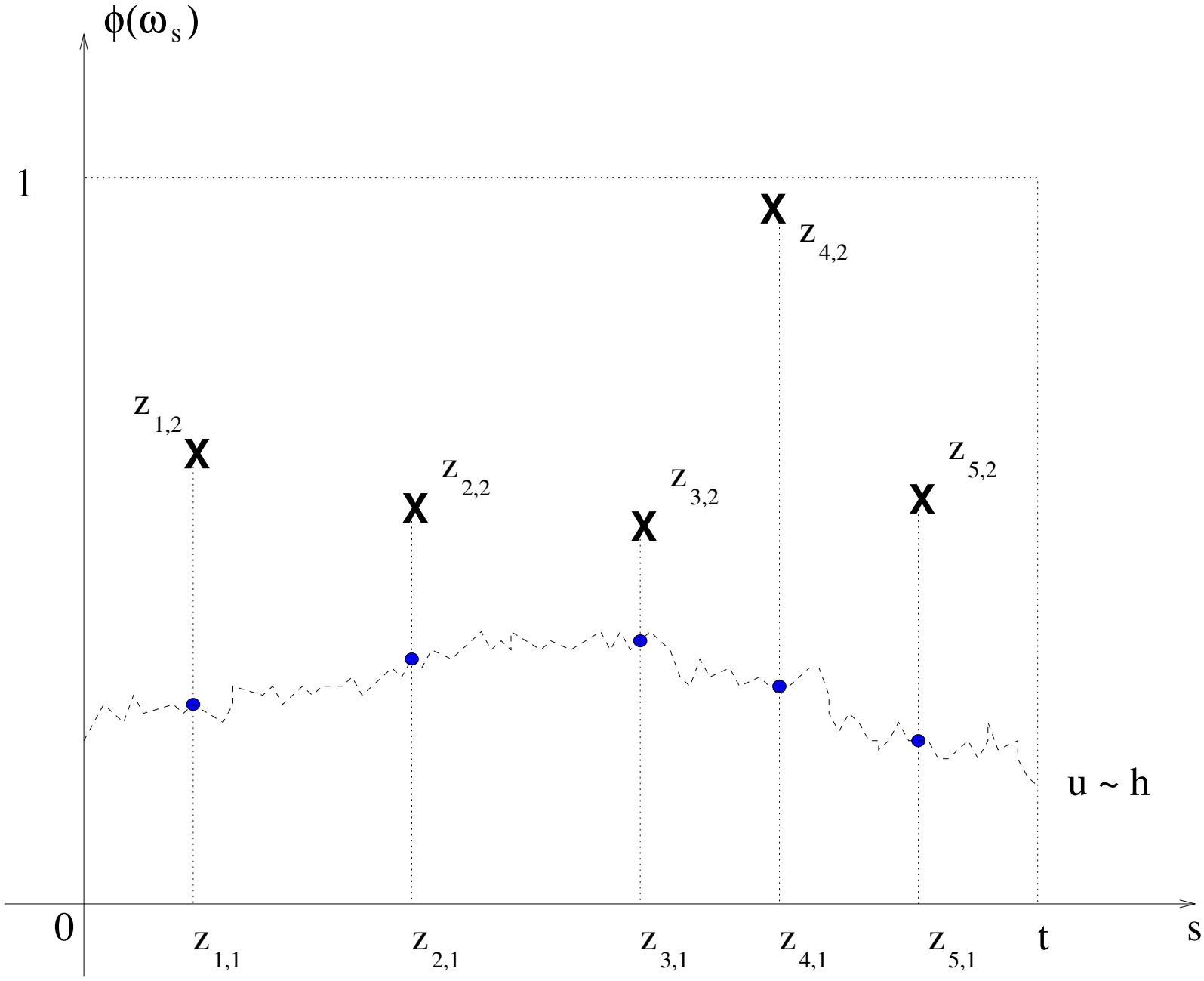
Left with carrying out step 2 of Algorithm in 76, and dealing with infinite paths. Ideas?

Event of equal probability: connection with Poisson process

Theorem

Let Φ be a homogeneous Poisson process of intensity r on $[0,t] \times [0,1]$ and N is the number of points of Φ below the graph $s \mapsto \phi(\omega_s)$, $s \in [0,t]$, then:

$$\mathrm{P}\left[\, N = 0 \,|\, \omega \,\right] = \exp\left\{ -r \int_0^t \phi(\omega_s) \mathrm{d}s \right\}.$$


$\phi(\omega_s)$
1
0
$z_{1,2}$
$z_{2,2}$
$z_{3,2}$
$z_{4,2}$
$z_{5,2}$
$z_{1,1}$
$z_{2,1}$
$z_{3,1}$
$z_{4,1}$
$z_{5,1}$
t
s
u ~ h

Idealized rejection sampler

1. Simulate $\omega \sim \mathbb{Z}^{(t,x)}$
2. Simulate a $Po(r)$ process $\Phi = \{z_1, z_2, \ldots, z_\kappa\}$, $z_j = (z_{j,1}, z_{j,2})$, $z_{j,1} \in [0,t], z_{j,2} \in [0,1]$, $1 \leq j \leq \kappa$
3. Compute the acceptance indicator I:

$$I := \prod_{j=1}^{\kappa} \mathbb{I}\left[\phi(\omega_{z_{j,1}}) < z_{j,2}\right]$$

4. If $I = 1$ accept ω, otherwise return to 1 and retry.

Retrospective Exact Simulation

EA1

1. Simulate $\Phi = \{z_1, z_2, \ldots, z_\kappa\}$
2. Simulate $u \sim h(u)$ and the values of $\omega \sim \mathbb{W}^{(t,x,u)}$, at the time instances $z_{j,1}$, $1 \leq j \leq \kappa$, therefore:

$$S(\omega) = \{(0, x), (z_{1,1}, \omega_{z_{1,1}}), \ldots, (z_{\kappa,1}, \omega_{z_{\kappa,1}}), (t, u)\}$$

3. Compute the acceptance indicator I.
4. If $I = 1$ then accept and return the proposed skeleton $S(\omega)$; otherwise return to 1 and retry.

Implications of EA

At first instance, we obtain an **exact algorithm** for simulation of V_t given V_0, at least for diffusions with certain conditions on their coefficients.

Obtaining values in $[0, t]$?

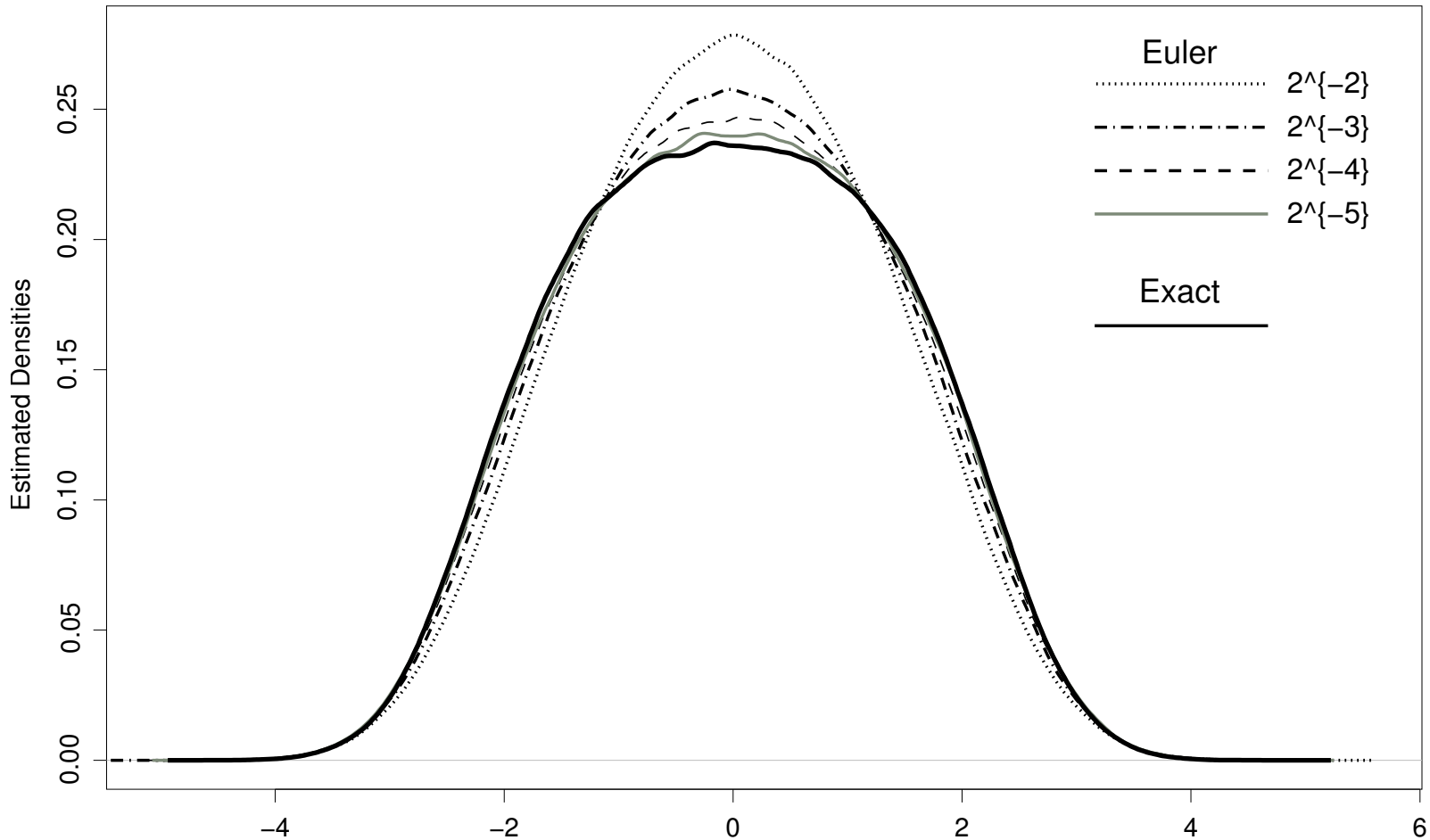
In fact, it opens the way to a completely different perspective on simulation and inference for stochastic processes

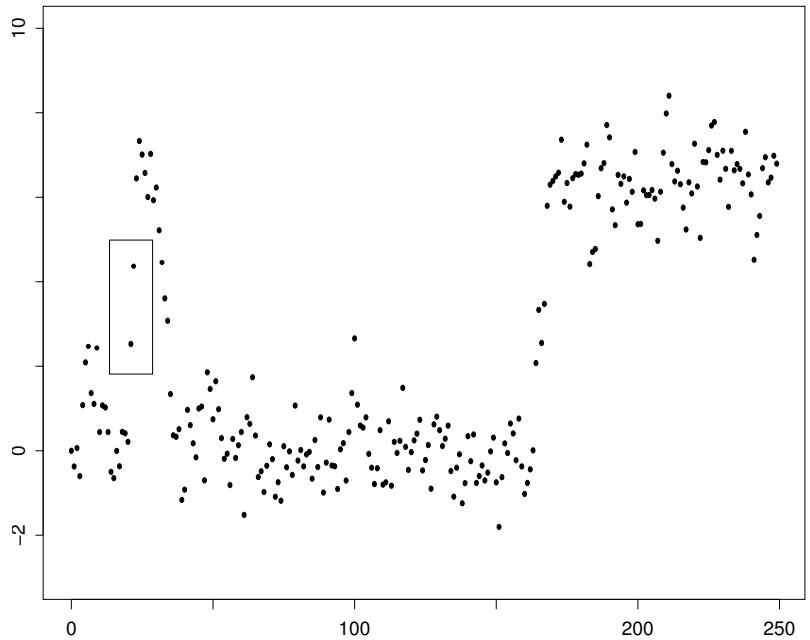
Exact simulation of other path functionals (e.g maximum, hitting times, etc)

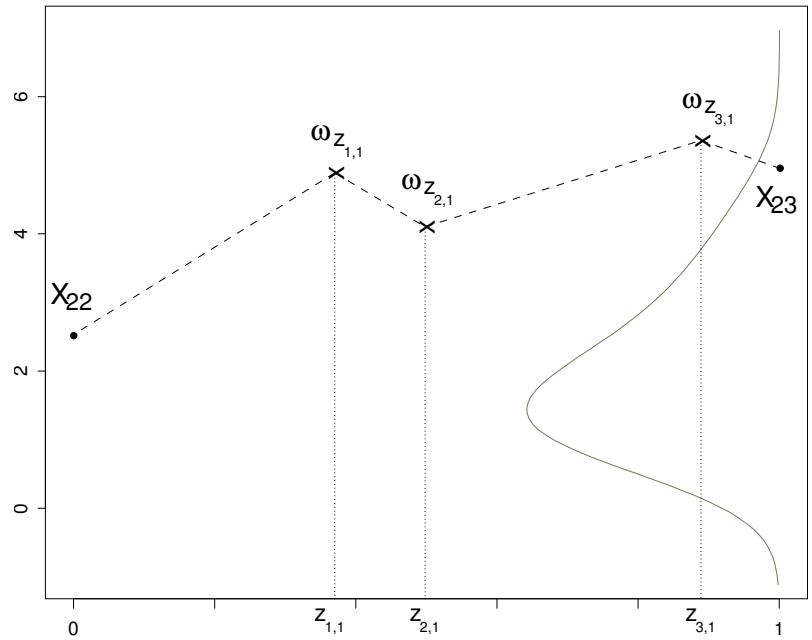
Simple example

$$\mathrm{d}X_s = \sin(X_s)\mathrm{d}s + \mathrm{d}B_s$$

This diffusion is not analytically tractable. Clearly, X in the class, $I = -1/2$ and $r = 9/8$. Exact Simulation of X_1:




10
0
−2
0
100
200
250


6
4
2
0
$\omega_{z_{1,1}}$
$\omega_{z_{2,1}}$
$\omega_{z_{3,1}}$
X_{22}
X_{23}
0
$z_{1,1}$
$z_{2,1}$
$z_{3,1}$
1

Final remark about the diffusion transformation

Critical to the procedure described is the Lamperti transformation (32) which transform V into a process with constant diffusion coefficient. This transformation has impact on ALL MC procedures for diffusions (see later).

For 1-d diffusions it exists under mild conditions on σ. For multi-d processes it might be intractable, or it might even not exist. The class of processes for which the transformation exists is known as reducible diffusions. We will return to this point later in the course.

Preamble: Some computational elements

- Generic Importance Sampling
- Brief introduction to MCMC
- Gibbs sampler and Data Augmentation
- Efficiency of Data Augmentation

Or jump to 116

Importance sampling and identities

Importance sampling (IS) is a classic Monte Carlo technique for obtaining samples from a probability measure $\mathbb{P}$ using samples from another probability measure $\mathbb{Q}$, see for example Chapter 2.5 of [Liu, 2008] for an introduction. Mathematically it is based on the concept of *change of measure*.

Suppose that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ with Radon-Nikodym density $f(x) = \mathbb{P}(\mathrm{d}x)/\mathbb{Q}(\mathrm{d}x)$. Then, in its simplest form IS consists of constructing a set of weighted particles (x_i, w_i), $i = 1, \ldots, N$, where $x_i \sim \mathbb{Q}$, and $w_i = f(x_i)$. This set gives a Monte Carlo approximation of $\mathbb{P}$, in the sense that for suitably integrable functions g, we have that

$$\frac{\sum_{i=1}^{N} g(x_i) w_i}{N} . \tag{34}$$

is an unbiased and consistent estimator of

$$\mathbb{E}_{\mathbb{P}}[g] := \int g(x) \mathbb{P}(\mathrm{d}x) .$$

IS can be cast in much more general terms, an extension particularly attractive in the context of stochastic processes. First, note that in most applications f is known only up to a normalising constant, $f(x) = cf_u(x)$, where only f_u can be evaluated and

$$c = \mathbb{E}_{\mathbb{Q}}[f_u]\,. \tag{35}$$

The notion of a properly weighted sample refers to a set of weighted particles (x_i, w_i), where $x_i \sim \mathbb{Q}$ and w_i is an unbiased estimator of $f_u(x_i)$, that is

$$\mathbb{E}_{\mathbb{Q}}[w_i \mid x_i] = f_u(x_i)\,.$$

Then for any integrable g

$$\mathbb{E}_{\mathbb{Q}}[gw] = \mathbb{E}_{\mathbb{P}}[g]\,\mathbb{E}_{\mathbb{Q}}[w]\,. \tag{36}$$

Rearranging the expression we find that a consistent estimator of $\mathbb{E}_{\mathbb{P}}[g]$ is given by

$$\frac{\sum_{i=1}^{N} g(x_i) w_i}{\sum_{i=1}^{N} w_i} . \tag{37}$$

When w_i is an unbiased estimator of $f(x_i)$ we have the option of using (34), thus yielding an unbiased estimator. However, (37) is a feasible estimator when c is unknown.

(37) is consistent and under moment conditions its asyptotic variance is

$$\frac{1}{N} Var\left(f(g - \mathbb{E}_{\mathbb{P}}[g])\right)$$

which should compare with the exact variance of (34)

$$\frac{1}{N} Var(fg)$$

IS includes exact simulation as a special case when $\mathbb{Q} = \mathbb{P}$. Another special case is rejection sampling (RS), which assumes further that $f_u(x)$ is bounded in x by some calculable $K < \infty$. Then, if we accept each draw x_i with probability $f_u(x_i)/K$, the resulting sample (of random size) consists of independent draws from $\mathbb{P}$. This is a special case of the generalised IS where w_i is a binary 0-1 random variable taking the value 1 with probability $f_u(x_i)/K$.

The main identity:

$$c = \mathbb{E}_{\mathbb{Q}}[w] \tag{38}$$

Brief intro to MCMC

Aim is again to sample from a probability measure $\mathbb{P}$ on some state-space $(\Omega, \mathcal{F})$. IS and its variations are global sampling methods. Thus they might be very inefficient if $\mathbb{Q}$ is different from $\mathbb{P}$. In any case, they will be increasingly worse (often exponentially so) when the dimension of the state space grows. One approach is to try to apply them sequentially, thus yielding sequential MC (SMC) methods, e.g particle filters

Another direction is to resort to iterative local algorithms. Markov chain Monte Carlo is the class of such methods. The samples are only asymptotically (in the number of iterations) drawn from $\mathbb{P}$ and are correlated (latter is true also for IS).

Examples of $\mathbb{P}$:

- Posterior distribution of a high-dimensional parameter vector in Bayesian statistics: $\mathbb{P}(\mathrm{d}\theta) = L(Y|\theta)\mathbb{Q}(\mathrm{d}\theta)$ (e.g the unobserved values of a spatial process observed with noise and hyperparameters controlling spatial correlation). Obtain samples to carry out statistical inference
- The uniform distribution on a constrained space (e.g counts on cells of a contingency table with fixed margins). Counting, MC tests
- The law of a diffusion conditioned to its endpoints

Assume for simplicity that $\mathbb{P}(\mathrm{d}\omega) = \pi(\omega)\mathbb{Q}(\mathrm{d}\omega)$, although this is not necessary to define the dynamics of MH, see e.g [Tierney, 1998].

The main idea is to produce a Markov chain which has transition kernel $P(\omega, \mathrm{d}\phi)$ invariant with respect to π and which has $\mathbb{P}$ as its unique limiting distribution, i.e it is ergodic. By invariance we mean:

$$\mathbb{P}(\mathrm{d}\phi) = \int_{\Omega} \pi(\omega) P(\omega, \mathrm{d}\phi) \mathbb{Q}(\mathrm{d}\omega)$$

When π is invariant for P, then under relatively mild conditions the chain will be ergodic. It is a necessary condition for ergodicity

Metropolis-Hastings algorithm

Let $q(\omega, \phi)$ denote a probability density function (in ϕ, w.r.t $\mathbb{Q}$) on Ω; this is called the proposal density. Choose ω_0 and then for each $n \geq 0$,

1. given ω_n propose $\phi_{n+1} \sim q(\omega_n, \cdot)$;
2. calculate the acceptance ratio

$$\alpha(\omega_n, \phi_{n+1}) = \frac{q(\phi_{n+1}, \omega_n)\pi(\phi_{n+1})}{q(\omega_n, \phi_{n+1})\pi(\omega_n)} \ ;$$

3. accept ϕ_{n+1}, setting $\omega_{n+1} = \phi_{n+1}$ with probability $1 \wedge \alpha$;
4. otherwise just set $\omega_{n+1} = \omega_n$.

Note that again, the normalizing constant of π is not necessary to carry out the algorithm. Examples of q: $q(\omega, \phi) = q(\phi)$ (independence sampler), $q(\omega, \phi) \propto \exp\{-||\phi - \omega||^2/(2\sigma^2)\}$ (random walk Metropolis, in Euclidean spaces)

Note that the transition kernel P has density p given by

$$p(\omega, \phi) = q(\omega, \phi)\alpha(\omega, \phi)\,, \quad \omega \neq \phi \quad \text{w.r.t } \mathbb{Q}$$

and probability of remaining at the same point

$$r(\omega) = \int_{\Omega} q(\omega, \phi)(1 - \alpha(\omega, \phi))\mathbb{Q}(\mathrm{d}\phi)$$

(assuming that, as will be in our examples, that the probability to propose exactly ω is 0)

We show that the MH kernel is invariant w.r.t $\mathbb{P}$. In fact, is satisfies a stronger property, that of that it is reversibility w.r.t $\mathbb{P}$, a property which implies invariance, and according to which the joint distribution of successive values is the same regardless of time-ordering:

$$\mathbb{P}(\mathrm{d}\omega)P(\omega, \mathrm{d}\phi) = \mathbb{P}(\mathrm{d}\omega)P(\omega \mathrm{d}\phi)$$

To see that this implies invariance, integrate both sides over ϕ.

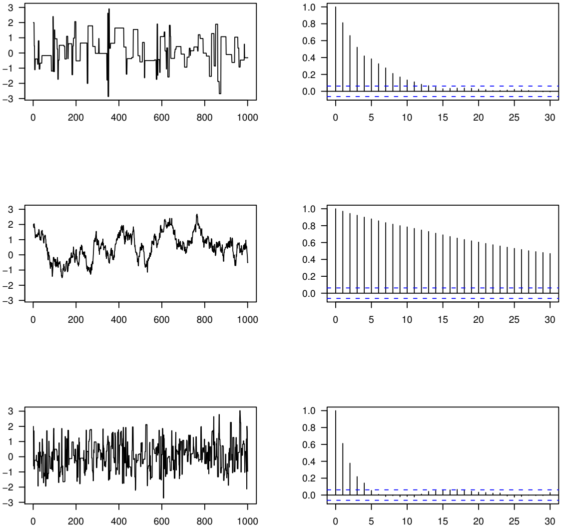
Note that it suffices to check this condition for $\omega \neq \phi$, in which case we need to check that the densities are in detailed balance

$$\pi(\omega)p(\omega, \phi) = \pi(\phi)p(\phi, \omega) \quad \omega \neq \phi$$

The LHS is given by:

$$\pi(\omega)q(\omega, \phi) \min\left\{1, \frac{q(\phi, \omega)\pi(\phi)}{q(\omega, \phi)\pi(\omega)}\right\} = \min\{\pi(\omega)q(\omega, \phi), q(\phi, \omega)\pi(\phi)\}$$

Scaling MH



Gibbs sampler

Another very popular (in particular in statistics) variant of MCMC is the Gibbs Sampler (GS). This presupposes a decomposition of ω into d components $\omega = (\omega^1, \ldots, \omega^d)$ and the existence of the corresponding conditional densities $\pi(\omega^i \mid \{\omega^j, j \neq i\})$

Random scan Gibbs sampler: iterate the following

1. choose I from $U(\{1, 2, \ldots d\})$
2. Replace ω^I by a random draw from $\pi(\omega^I | \{\omega^j, j \neq I\})$.

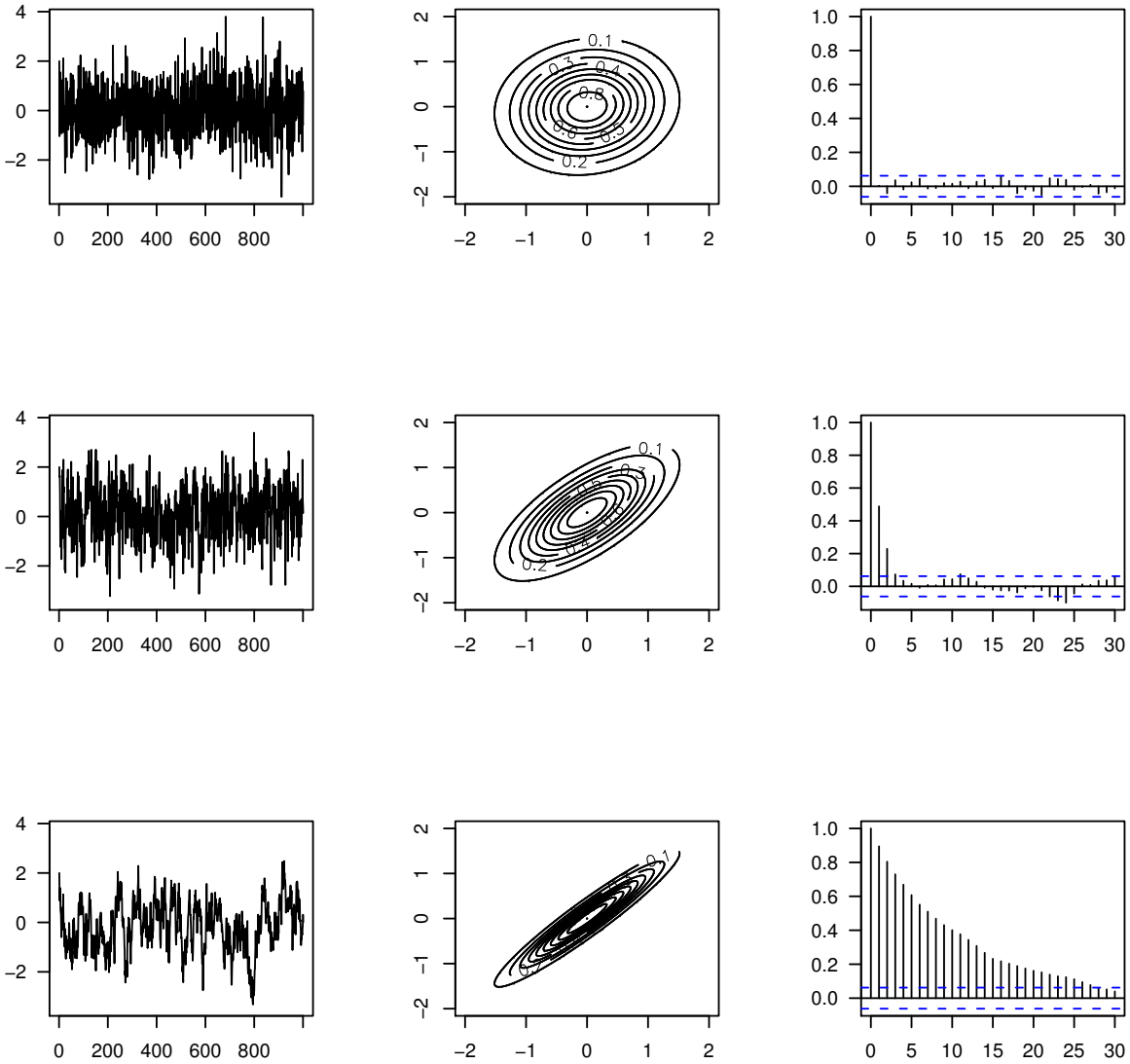
Deterministic scan Gibbs sampler, DUGS: instead of chosing a random component to update, systematically update each component in turn.

In general (unless $d = 2$) this generates a Markov chain invariant but not reversible w.r.t π. Note however that each conditional step is reversible w.r.t π.

The decomposition of ω (and π) into d components is driven by the following two conflicting aims:

- Be able to simulate directly/efficienty the variables within each block j
- The variables corresponding to different blocks are weakly dependent

The role of dependence



Simplifying individual steps: Metropolis-within-Gibbs

In many intresting non-trivial examples it will not be possible to carry out directly the conditional simulation in some or all steps of the Gibbs sampler (where the groups have been chosen mainly with view to aim 2 mentioned before)

Nevertheless, when updating component i we can apply perform a MH step which targets $\pi(\omega^i|\{\omega^j, j \neq i\})$, or apply any other kernel which leaves this conditional invariant

Data augmentation

The case $d = 2$ is quite special, since it is easier to study, but it also naturally appears in many application, in particular to the so-called missing data problems, which we will study in more detail.

Sometimes the GS in this case is called the Data Augmentation algorithm, from the historical developement of a relevant algorithm for inference in missing data problems. Close links to EM.

For this case one can also get an interesting characterization of the rate of convergence of the algorithm, which explains the previous plots

Rate of convergence of DA

$$\gamma = 1 - \inf_{h \in L^2(\pi)} \frac{\mathbb{E}(\mathrm{Var}(h(\omega_1)|\omega_2))}{\mathrm{Var}(h(\omega_1))}$$

Then γ is known as the Bayesian fraction of missing information (see for example [Rubin, 2004, Meng and van Dyk, 1997]).

γ **is also the rate of convergence of DA**, which practically means that in stationarity the algorithm needs $-1/\log\gamma$ time to mix around the state space (forget the initial value). Practically it also means that the effective sample size is about $(1-\gamma)/(1+\gamma)$

We will revisit DA in the context of partially observed stochastic processes

Part IV:MC-based likelihood inference for discretely-observed diffusions with known constant diffusivity

- Missing data problems and formal data augmentation (DA)
- DA for discretely-observed diffusions with known constant diffusion coefficient
- Conditional distribution of the missing data: diffusion bridges
- Likelihood ratios for diffusion bridges, transition density identities, connections to literature
- An MCMC scheme for parameter estimation

To avoid excessive notation we focus on time-homogeneous diffusions, although this is only for convenience