1 Convergence of the MLE for the drift

Let us consider again the Ornstein–Uhlenbeck equation:

$$\mathrm{d}X_t = -\alpha\, X_t\, \mathrm{d}t + \mathrm{d}W_t, \qquad X_0 = x_0, \qquad \alpha < 0. \tag{1}$$

For simplicity, we will consider only the case of a deterministic initial condition. We begin this week's lectures by proving an asymptotic result for the maximum likelihood estimator (MLE) that we derived last week for the drift coefficient:

$$\hat{\alpha} = -\frac{\int_0^T X_t\, \mathrm{d}X_t}{\int_0^T |X_t|^2\, \mathrm{d}t}.$$

Employing (1) to formally rewrite $\mathrm{d}X_t$ as $-\alpha\, X_t\, \mathrm{d}t + \mathrm{d}W_t$, we obtain

$$\hat{\alpha} = \alpha - \frac{\int_0^T X_t\, \mathrm{d}W_t}{\int_0^T |X_t|^2\, \mathrm{d}t}.$$

The rigorous justification of this step would require the rigorous definition of the stochastic integral with respect to X_t, which we will not do here. Remember, however, that we observed a similar equivalence at the finite-dimensional level when we calculated the probability of ruin of a gambler in the first problem sheet.

The exact solution to (1) is given by

$$X_t = \mathrm{e}^{-\alpha t}\, x_0 + \int_0^t \mathrm{e}^{-\alpha(t-s)}\, \mathrm{d}W_s. \tag{2}$$

As you showed in the coursework, the law of the stochastic integral in this equation is

$$\int_0^t \mathrm{e}^{-\alpha(t-s)}\, \mathrm{d}W_s \sim \mathcal{N}\left(0, \int_0^t \mathrm{e}^{-2\alpha(t-s)}\, \mathrm{d}s\right) = \mathcal{N}\left(0, \frac{1-\mathrm{e}^{-2\alpha t}}{2\alpha}\right).$$

We deduce that X_t is a Gaussian process and, using the assertion (3) in Slutsky's theorem, that $X_t \to \mathcal{N}\left(0, \frac{1}{2\alpha}\right)$ in distribution as $t \to \infty$,

Theorem 1 (Slutsky)**.** *Assume that $X_n \to X$ in distribution and $Y_n \to c$ in distribution as $n \to \infty$, for a random variable X and a constant c. Then*

$$X_n + Y_n \to X + c \quad \text{in distribution as } n \to \infty. \tag{3}$$

and

$$\frac{X_n}{Y_n} \to \frac{X}{c} \quad \text{in distribution as } n \to \infty.$$

Remark 1. You may be wondering whether (3) can be generalized to the case where $Y_n \to Y$ in distribution for some non-constant random variable Y. In general, the answer is no: consider for example the sequences $X_n = Z$, $Y_n = -Z$, for a random variable $Z \sim \mathcal{N}(0,1)$. It is clear that $X_n \to Z$ and $Y_n \to Z$ in distribution, but $S_n := X_n + Y_n = 0$ converges to 0 in distribution. ⊘

The probability measure measure associated with $\mathcal{N}\left(0, \frac{1}{2\alpha}\right)$ is called the *invariant measure* of the stochastic differential equation, and we will denote its density by $\rho_\infty(x)$. From the expression

of the exact solution (2), we can also obtain the following bounds on the moments of X_t, which we will require for proving Lemma 3 and Theorem 4.

Lemma 2 (Uniform-in-time bound on the moments of X_t)**.** *Let X_t be given by* (2)*. Then for every $m \in \mathbb{N}$ there exists a constant C such that*

$$\mathbb{E}|X_t|^m \leq C \qquad \forall t \geq 0.$$

Proof. From Jensen's inequality,

$$\left(\frac{u}{2} + \frac{v}{2}\right)^m \leq \frac{u^m}{2} + \frac{v^m}{2} \qquad \forall u, v \in \mathbb{R}.$$

Employing this in (2), we obtain

$$\mathbb{E}|X_t|^m = 2^{m-1} \left|\mathrm{e}^{-\alpha t} x_0\right|^m + 2^{m-1} \mathbb{E}\left| \int_0^t \mathrm{e}^{-\alpha(t-s)} \, \mathrm{d}W_s \right|^m .$$

Since the stochastic integral is $\sim \mathcal{N}\left(0, \frac{1-\mathrm{e}^{-2\alpha t}}{2\alpha}\right)$, we can employ the formula for the moments of the normal distribution to deduce

$$\mathbb{E}|X_t|^m \leq 2^{m-1} \left|\mathrm{e}^{-\alpha t} x_0\right|^m + 2^{m-1} \sqrt{\frac{1-\mathrm{e}^{-2\alpha t}}{2\alpha}} (m-1)!!,$$

which concludes the proof. □

The last ingredient we need, in order to prove the convergence of the MLE estimator, is an *ergodicity result*. Roughly speaking, ergodicity means that time averages of an observable $f(X_t)$ converge to space averages – averages with respect to the invariant measure of the SDE.

Lemma 3 (Ergodicity for the Ornstein–Uhlenbeck process)**.** *Let $f(x) = x^2$ and X_t be the solution to* (1)*. Then there exists a constant C that does not depend on T such that*

$$\mathbb{E}\left| \frac{1}{T} \int_0^T f(X_t) \, \mathrm{d}t - \int_{\mathbb{R}} f(x) \, \rho_\infty(x) \, \mathrm{d}x \right|^2 \leq \frac{C}{T} \qquad \forall T > 0. \tag{4}$$

Remark 2. Here we prove this result only for the function $f(x) = x^2$ and for the Ornstein–Uhlenbeck process, but the statement holds more generally. For example, it should appear clearly from the proof below that (4) also holds for $f(x) = x^3$ (with a different constant C). In more general contexts, the crux of the problem lies in proving that (5) admits a solution and that this solution does not grow too fast as $x \to \infty$. ⊘

Proof. Let $\mathcal{L}$ be the generator of (1) and consider the following partial differential equation, known as a *Poisson equation*:

$$-\mathcal{L}\phi(x) = f(x) - \mu_f, \qquad \mathcal{L} = -\alpha \, x \, \partial_x + \frac{1}{2}\partial_x^2, \qquad \mu_f := \int_{\mathbb{R}} f(x) \, \rho_\infty(x) \, \mathrm{d}x. \tag{5}$$

If $\mathcal{L}$ and f were the generator of a general SDE and a general observable, respectively, showing the existence of a solution to this equation would be quite difficult. In our simple setting, however, an explicit solution can be obtained simply from the ansatz $\phi(x) = ax^2 + b\,x + c$: substituting

in (5), we find $\phi(x) = \frac{x^2}{2\alpha} + c$ for any constant c. Since the value of c has no impact on the forthcoming calculations, we take $c = 0$ for simplicity. Employing Itô's formula for $Y_t = \phi(X_t)$, we obtain

$$\phi(X_T) - \phi(X_0) = \int_0^T f(X_t) - \mu_f \,\mathrm{d}t + \int_0^T \phi'(X_t)\,\mathrm{d}W_t.$$

It follows from the standard inequality $|u+v|^2 \leq 2u^2 + 2v^2$ that

$$\begin{aligned}\mathbb{E}\left|\frac{1}{T}\int_0^T f(X_t) - \mu_f \,\mathrm{d}t\right|^2 &\leq \frac{2}{T^2}\,\mathbb{E}|\phi(X_T) - \phi(X_0)|^2 + \frac{2}{T^2}\,\mathbb{E}\left|\int_0^T \phi'(X_t)\,\mathrm{d}W_t\right|^2 \\ &\leq \frac{2}{T^2}\,\mathbb{E}|\phi(X_T) - \phi(X_0)|^2 + \frac{2}{T^2}\int_0^T \mathbb{E}|\phi'(X_t)|^2 \mathrm{d}s.\end{aligned}$$

Since ϕ a polynomial of degree 2, and since all the moments of X_t are bounded uniformly on the interval $[0,\infty)$ by Lemma 2, we deduce

$$\mathbb{E}\left|\frac{1}{T}\int_0^T f(X_t) - \mu_f \,\mathrm{d}t\right|^2 \leq \frac{C}{T}.$$

which concludes the proof. □

Theorem 4 (Convergence of the MLE)**.** *For all $0 \leq \beta < 1/2$, It holds that*

$$T^\beta(\hat{\alpha} - \alpha) \to 0 \quad \textit{in distribution as } T \to \infty.$$

Remark 3. In fact, it is possible to show that $\sqrt{T}(\hat{\alpha} - \alpha) \to \mathcal{N}(0, 2\alpha)$ in distribution, which is a stronger statement but requires more machinery. ⊘

Proof. Employing Itô's formula, we notice that

$$\frac{1}{2}(X_T^2 - X_0^2) = \int_0^T \left(-\alpha X_t^2 + \frac{1}{2}\right)\,\mathrm{d}t + \int_0^T X_t\,\mathrm{d}W_t,$$

so we can rewrite the formula for $\hat{\alpha}$ as

$$\hat{\alpha} = -\frac{X_T^2 - X_0^2 - T}{2\int_0^T |X_t|^2\,\mathrm{d}t} = \frac{1}{\frac{2}{T}\int_0^T |X_t|^2\,\mathrm{d}t} - \frac{\frac{1}{T}(X_T^2 - X_0^2)}{\frac{2}{T}\int_0^T |X_t|^2\,\mathrm{d}t}.$$

It follows that

$$T^\beta(\alpha - \hat{\alpha}) = \frac{\alpha\, T^\beta \frac{2}{T}\int_0^T |X_t|^2 - \frac{1}{2\alpha}\,\mathrm{d}t}{\frac{2}{T}\int_0^T |X_t|^2\,\mathrm{d}t} + \frac{T^{\beta-1}(X_T^2 - X_0^2)}{\frac{2}{T}\int_0^T |X_t|^2\,\mathrm{d}t} =: \frac{N_1}{D} + \frac{N_2}{D}.$$

We now show that $N_1 \to 0$, $N_2 \to 0$ and $D \to \frac{1}{2\alpha}$ in distribution as $T \to \infty$, which by repeated application of Sultsky's theorem will conclude the proof. The convergence of N_1 and D follows from Lemma 3 and the fact that convergence in $L^2(\Omega)$ implies convergence in distribution. For N_2, notice that

$$\mathbb{E}|N_2| \leq T^{\beta-1}(\mathbb{E}|X_T|^2 + \mathbb{E}|X_0|^2),$$

which, by Lemma 2, implies that $N_2 \to 0$ in $L^1(\Omega)$ and therefore also in distribution. □

6.3. Transformation methods for MLE.

In this section we will justify the transformation that allowed us to use an additive SDE

$$dX_t = b(X_t;\theta)\,dt + dW_t$$

for the MLE instead of the original SDE

$$dX_t = b(X_t;\theta)\,dt + \sigma(X_t;\theta)\,dW_t.$$

We note that the drift in both equations might not be the same.

6.3.1. Lamperti's transformation

For SDEs in one dimension, it is possible to map multiplicative noise to additive noise.
If we have an Itô SDE with multiplicative noise

$$dX_t = b(X_t;\theta)\,dt + \sigma(X_t;\theta)\,dW_t$$

and ask whether we can find a transformation $Z_t = h(X_t)$ that maps this SDE into one with additive noise.
We apply Itô's formula and obtain

$$dZ_t = \mathcal{L}h(X_t)\,dt + h'(X_t)\,\sigma(X_t)\,dW_t$$

where $\mathcal{L}f(x) = \frac{df}{dx}b(x) + \frac{d^2f}{dx^2}\frac{\sigma^2(x)}{2}$.

In order to obtain an SDE for Z_t with additive noise and diffusion coefficient equal to 1, we need to impose

$$h'(x)\,\sigma(x) = 1$$

which implies

$$\boxed{h(x) = \int_{x_0}^{x} \frac{1}{\sigma(x)}\,dx}.$$

Now we have that

$$\mathcal{L}h(x) = \frac{dh}{dx}b(x) + \frac{d^2h}{dx^2}\frac{\sigma^2(x)}{2}$$
$$= \frac{1}{\sigma(x)}b(x) - \frac{\sigma'(x)}{\sigma^2(x)}\frac{\sigma^2(x)}{2}$$
$$= \frac{b(x)}{\sigma(x)} - \frac{\sigma'(x)}{2}.$$

Since $Z_t = h(X_t)$, $X_t = h^{-1}(Z_t)$ and $b(X_t) = b(h^{-1}(Z_t))$, etc.

So the SDE for Z_t is

$$\boxed{dZ_t = \left(\frac{b(h^{-1}(Z_t))}{\sigma(h^{-1}(Z_t))} - \frac{\sigma'(h^{-1}(Z_t))}{2}\right)dt + dW_t}$$

where $h(x) = \int_{x_0}^{x} 1/\sigma(x)\,dx$.

This is called the Lamperti transformation.

example: Consider the Cox-Ingersoll-Ross (CIR) SDE:

$$dX_t = (\mu - \alpha X_t)dt + \sigma\sqrt{X_t}\,dW_t, \quad X_0 = x > 0.$$

Using the formula for $h(x)$ with $x_0 = 0$, we can easily deduce that $\boxed{h(x) = \frac{2\sqrt{x}}{\sigma}}$

the generator $\mathcal{L}$ of the CIR process is

$$\mathcal{L}f(x) = (\mu - \alpha x)\frac{df}{dx} + \frac{\sigma^2}{2}x\frac{d^2f}{dx^2}.$$

So, when applied to h we have

$$\mathcal{L}h(x) = (\mu - \alpha x)\frac{1}{\sigma\sqrt{x}} + \frac{\sigma^2 x}{2}\left(-\frac{1}{2\sigma x^{3/2}}\right)$$

$$= \frac{\mu}{\sigma\sqrt{x}} - \frac{\alpha\sqrt{x}}{\sigma} - \frac{\sigma}{4\sqrt{x}}$$

$$= \left(\frac{\mu}{\sigma} - \frac{\sigma}{4}\right)x^{-1/2} - \frac{\alpha}{\sigma}x^{1/2}.$$

So, for $Y_t = 2\sqrt{X_t}/\sigma$, the CIR SDE becomes

$$dY_t = \left(\frac{\mu}{\sigma} - \frac{\sigma}{4}\right)\frac{1}{\sqrt{X_t}}dt - \frac{\alpha}{\sigma}\sqrt{X_t}\,dt + dW_t$$

or

$$\boxed{dY_t = \left[\left(\frac{2\mu}{\sigma^2} - \frac{1}{2}\right)\frac{1}{Y_t} - \frac{\alpha}{2}Y_t\right]dt + dW_t}$$

Note that if $\mu = \frac{\sigma^2}{4}$, the above equation becomes the SDE for the Ornstein-Uhlenbeck process!

Note that:

- this transformation is really useful when estimating parameters using MLE, since it allows us to obtain additive SDEs with diffusion coefficient 1.
- Such a transformation does not exist in higher dimensions. In fact it is not possible, in general, to transform a multi-dimensional multiplicative SDE into an additive one.
- If it is possible, the process is called reducible. You can find conditions on coefficients so that SDE is reducible in Ait-Sahalia, The Annals of Statistics 36(2) 906-937 (2008).

6.5. Bayesian estimation

In certain applications, it is important to understand the uncertainty in the parameter θ which we are estimating. In these cases, rather than obtaining a single "best" estimate for θ, it is far more informative to obtain a distribution of possible values that θ might take given the observed data. One way of obtaining this is to use Bayes' rule, which incorporates data as well as prior information about the parameters in order to obtain a distribution of the possible values of θ. It states that

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(y|\theta)\,\pi_0(\theta)}{\mathbb{P}(y)}.$$

where $\mathbb{P}(y|\theta) = L(y|\theta)$ is the likelihood function (which we have seen before), and $\pi_0(\theta)$ is a prior distribution on the unknown parameter θ – it incorporates information that we have already learned on θ.
$\mathbb{P}(\theta|y)$ is called the posterior density.

In this approach, we see θ as a random variable itself, and its distribution $\mathbb{P}(\theta|y)$ characterises our certainty or uncertainty about the real value of θ, given what we have observed.

The main complication arises from the computation of the denominator

$$\mathbb{P}(y) = \int_{\Theta} \mathbb{P}(y|\theta)\,\pi_0(\theta)\,d\theta$$

which can be a high dimensional integral – hard to compute.

There are many ways to deal with this denominator. A natural strategy would be to choose a prior distribution $\pi_0(\theta)$ in such a way that $\mathbb{P}(y)$ is known although this might not always be possible.

Another alternative is to "compromise" between MLE and Bayesian inference: using a <u>Maximum a posteriori estimator</u>.
This method consists on <u>estimating θ using the mode</u> (rather than the mean) of the posterior distribution. Since the mode <u>does not depend</u> on a normalisation constant, we do not need to compute $\mathbb{P}(y)$!
The maximum a posteriori (MAP) estimator is then defined as follows:

$$\hat{\theta}_{MAP}(y) = \arg\max_{\theta \in \Theta} \mathbb{P}(\theta|y)$$

$$= \arg\max_{\theta \in \Theta} \frac{\mathbb{P}(y|\theta)\,\pi_0(\theta)}{\mathbb{P}(y)}$$

$$= \arg\max_{\theta \in \Theta} \mathbb{P}(y|\theta)\,\pi_0(\theta).$$

<u>Note that</u>: If the prior distribution is uniform (i.e., if $\pi_0(\theta) = \text{constant}$), the MAP estimator coincides with the MLE!

We can compute the MAP estimator <u>analytically</u> (if we know a closed form for the mode of the posterior distribution - this is the case, for example, when both the prior and the likelihood distributions are Gaussian), <u>numerically</u>, using optimisation tools, such as gradient descent

or Newton's method – Note that this requires first and/or second derivatives – or by using other techniques such as an expectation - maximisation algorithm or Monte Carlo simulations using simulated annealing

A final possibility is to use Markov Chain Monte Carlo — construct a Markov chain whose stationary distribution is $P(\theta|y)$.

There are many examples of application of MCMC to parameter estimation. The most well known is probably the Gibbs sampler, which is a special case of the Metropolis-Hastings algorithm.

The choice of prior plays an important role on the performance - certain algorithms can only use Gaussian priors, for example.

Finally, some of these methods can also be extended to estimate functions rather than parameters.

For more information/references about these methods you can always ask/email me!

5. Markov Chain Monte Carlo (MCMC)

In section 2 we discussed Monte Carlo methods to compute expectations of the form $\mathbb{E}(f(X))$, where $X \sim \pi$, a known distribution. This method relies on the assumption that it is possible and feasible to generate samples of π. this is not necessarily true in many scenarios (especially in higher dimensions). Therefore we need to develop methods to overcome this. We will first recall some definitions and properties of Markov chains / processes.

Definition: A discrete time stochastic process $\{X_n\}_{n \in \mathbb{N}}$, $X_n : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ is a Markov chain if

$$\mathbb{P}(X_{n+1} \in B \mid X_0, X_1, \dots, X_n) = \mathbb{P}(X_{n+1} \in B \mid X_n), \quad \forall B \in \mathcal{S}, n \in \mathbb{N}.$$

↳ Markov property

(The equivalent definition for continuous time stochastic process is (Markov process)

$$\mathbb{P}(X_t \in B \mid \mathcal{F}_s) = \mathbb{P}(X_t \in B \mid X_s), \quad \forall A \in \mathcal{S}, s, t \in T, s < t.$$

)

If the Markov chain starts at x, we use the notation

$$\mathbb{P}_x(X_n \in B) = \mathbb{P}(X_n \in B \mid X_0 = x).$$

We can write the finite dimensional distribution of a Markov chain with initial distribution $X_0 \sim \mu$ as follows:

$$\mathbb{P}_\mu(X_0 \in B_0, X_1 \in B_1, \dots, X_n \in B_n) =$$

$$= \int_{B_0} \mu \, d(y_0) \int_{B_1} \mathbb{P}(X_1 \in dy_1 \mid X_0 = y_0) \cdots \int_{B_n} \mathbb{P}(X_n \in dy_n \mid X_{n-1} = y_{n-1}).$$

Definition: A Markov chain $\{X_n\}_{n\in\mathbb{N}}$ is time-homogeneous if $\mathbb{P}(X_{n+1}\in B \mid X_n) = \mathbb{P}(X_1 \in B \mid X_0)$, $\forall n \geq 0$.

If a Markov chain is time homogeneous, we can write

$$\mathbb{P}(X_{n+1}\in B \mid X_n = x) =: p(x, B),$$

for some function $p: \mathcal{S}\times\mathcal{S} \longrightarrow [0,1]$, which is called a transition function. For fixed $x\in\mathcal{S}$, $p(x,\cdot)$ is a probability measure. Furthermore, $p(\cdot, B)$ is a measurable map.

If $\mathcal{S}$ is discrete (finite or countable), we can define a transition matrix $p = \{p(x,y),\ x,y\in\mathcal{S}\}$, where

$$\sum_{y\in\mathcal{S}} p(x,y) = 1, \quad p(x,y)\geq 0 \ \ \forall x,y\in\mathcal{S}$$

It is easy to see that

$$p(x,y) := \mathbb{P}(X_1 = y \mid X_0 = x) = \mathbb{P}_x(X_1 = y)$$

and

$$\mathbb{P}_\mu(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mu(x_0)\, p(x_0,x_1)\cdots p(x_{n-1},x_n).$$

Finally, for $n\geq 1$ we write $p^n(x,y) = \mathbb{P}_x(X_n = y)$.

An important consequence of the Markov property is the Chapman-Kolmogorov equation:

Theorem: Let X_n be a time-homogeneous Markov chain with discrete state space $\mathcal{S}$. Then, $\forall m, n \geq 0$

$$\mathbb{P}_x(X_{n+m} = y) = \sum_{z\in\mathcal{S}} \mathbb{P}_x(X_n = z)\, \mathbb{P}_z(X_m = y).$$

5.1. Stationary processes, stationary distributions and ergodicity.

An important property which we will use is stationarity.

Definition: We say that a Markov chain $\{X_n\}_{n\in\mathbb{N}}$ is stationary if, for every $n\in\mathbb{N}$, the joint distribution

$$\mathbb{P}(X_k\in B_0, X_{k+1}\in B_1, \dots, X_{k+n}\in B_n)$$

is independent of the offset $k>0$. In particular, from the $n=1$ case, we have

$$\mathbb{E}(X_i)=\mathbb{E}(X_0)=:I, \quad \forall i\in\mathbb{N},$$

for $n=2$

$$\mathrm{Var}(X_i)=\mathrm{Var}(X_0)=:\sigma^2, \quad \forall i\in\mathbb{N}$$

and

$$\mathrm{Cov}(X_i,X_j)=C(j-i), \quad \forall i,j\in\mathbb{N}.$$

Note that we can define this concept for continuous processes too. We saw this in section 3:
A stochastic process is (strictly) stationary if all its FDDs are invariant under time translation: for all $k\in\mathbb{N}$, $\forall t_i\in T$, $\{\Gamma_i\}_{i=1}^k\subset\mathcal{B}$,

$$\mathbb{P}(X_{t_1}\in\Gamma_1,\dots,X_{t_k}\in\Gamma_k)=\mathbb{P}(X_{s+t_1}\in\Gamma_1,\dots,X_{s+t_k}\in\Gamma_k)$$

for $s>0$: $s+t_i\in T$, $\forall i=1,\dots,k$. As for the discrete case, this implies that $\mathbb{E}(X_{t+s})=\mathbb{E}(X_t)$, $\forall s\in T$ and $C(t,s)=C(t-s)$. Furthermore, the law of X_t does not depend on t.

A stochastic process is wide sense stationary (WSS) (or second order, or weakly stationary) if

- $\mathbb{E}(X_t)$ does not depend on t
- $\mathrm{Cov}(X_t,X_s)$ is a function of $t-s$.

When we computed MC estimators in section 2, we considered r.v.s $\{X_n\}_{n\geq 0}$ which were identically distributed with distribution π and were independent. Now we are going to relax the independence assumption, i.e., we will use stationary Markov chains in order to allow for correlation between the variables of the chain at different times.

As before, given a stationary sequence Z_n, we can construct time-averages of the form

$$I_n = \frac{1}{n}\sum_{i=1}^{n} Z_i .$$

Then we have $\mathbb{E}(I_n) = I = \mathbb{E}(X_0)$, for all $n \geq 1$. So if we can generate realisations of the chain, I_n is an unbiased estimator for I!
What is its variance?

$$\mathrm{Var}\left(\sum_{i=1}^{n} Z_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(Z_i, Z_k) =$$
$$= \sum_{i=1}^{n} \mathrm{Var}(Z_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{Cov}(Z_i, Z_j).$$

If the Markov chain is stationary, $\mathrm{Var}(Z_i) = \mathrm{Var}(Z_1)$ and $\mathrm{Cov}(Z_n, Z_{n+k})$ is independent of n. Therefore

$$\mathrm{Var}\left(\sum_{i=1}^{n} Z_i\right) = n\,\mathrm{Var}(Z_1) + 2\sum_{k=1}^{n-1}(n-k)\,\mathrm{Cov}(Z_j, Z_{j+k})$$
$$= n\,\mathrm{Var}(Z_1) + 2\sum_{k=1}^{n-1}(n-k)\,C(k).$$

and we have

$$n\,\mathrm{Var}(I_n) = \sigma^2 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\,C(k).$$

Taking $n \to \infty$, and assuming that the limit

$$\lim_{n\to\infty}\left(\sigma^2 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}C(k)\right) = \sigma^2 + 2\sum_{k=1}^{\infty}C(k)$$

exists, then we might hope that there is a central limit theorem for I_n, i.e., that $\sqrt{n}(I_n - I)$ converges to a Gaussian distribution with mean 0 and variance $\sigma^2 + 2\sum C(k)$.

This is true, provided the distribution is also ergodic. We will now see what this means.

Definition: A probability measure π on $\mathcal{S}$ is a stationary distribution for the Markov chain X_n with transition matrix p if

$$\sum_x \pi(x)\, p(x,y) = \pi(y).$$

Equivalently, π is a stationary distribution for X_n if $X_n \sim \pi \Rightarrow X_{n+1} \sim \pi$.

Note that if π is a stationary distribution, then

$$\sum_x \pi(x)\, p^n(x,y) = \pi(y), \quad \forall n \geq 1.$$

If the state space $\mathcal{S}$ is finite dimensional, then a stationary distribution can be expressed as an N-dimensional vector (where N is the dimension of $\mathcal{S}$). Then if the chain has transition matrix p, π is a stationary distribution if and only if

$$p^* \pi = \pi.$$