4 Stochastic signal processing

Now that we have at least some idea of what stochastic signals are, we can begin to analyze the behavior of systems with stochastic inputs. Let's recall that a system is anything that transforms an input signal into an output signal. Let's keep things simple and work with continuous-time input and output signals, and consider the case of deterministic inputs first. A system $\mathcal{S}$ transforms an input signal $x : \mathbb{R} \to \mathbb{R}$ into an output signal $y : \mathbb{R} \to \mathbb{R}$, and we denote this fact by writing $y = \mathcal{S}[x]$. This notation signifies the fact that the input to $\mathcal{S}$ is the *entire* signal x, and the output of $\mathcal{S}$ is the *entire* signal y. Thus, the value $y(t) = \mathcal{S}[x](t)$ of the output at time t may, in principle, depend on all $x(s), s \in \mathbb{R}$.

Clearly, at this level of abstraction there is not a whole lot that can be done. So, it is useful to single out various types of systems:

- **causal** — when the current value of the output is not affected by the future values of the input. Formally, for any $t \in \mathbb{R}$ and for any two inputs x_1 and x_2, such that $x_1(s) = x_2(s)$ for all $s \leq t$, we have
$$\mathcal{S}[x_1](t) = \mathcal{S}[x_2](t).$$
- **memoryless** — when the current value of the output depends only on the current value of the input. Formally, for any $t \in \mathbb{R}$ and for any two inputs x_1 and x_2, such that $x_1(t) = x_2(t)$, we have
$$\mathcal{S}[x_1](t) = \mathcal{S}[x_2](t).$$
Any memoryless system is causal.
- **time-invariant** — when the output due to a time-shifted version of the input is the time-shifted version of the output. Formally, given a signal $v : \mathbb{R} \to \mathbb{R}$ and an arbitrary $\tau \in \mathbb{R}$, define its time shift $v_\tau : \mathbb{R} \to \mathbb{R}$ by $v_\tau(t) \triangleq v(t - \tau)$. Then
$$\mathcal{S}[x_\tau] = (\mathcal{S}[x])_\tau.$$
- **linear** — when the output due to a superposition of inputs is the superposition of the outputs. Formally, given any two input signals $x_1, x_2 : \mathbb{R} \to \mathbb{R}$ and any two real coefficients α_1, α_2,
$$\mathcal{S}[\alpha_1 x_1 + \alpha_2 x_2] = \alpha_1 \mathcal{S}[x_1] + \alpha_2 \mathcal{S}[x_2].$$

We have already seen how Markov chains can be viewed as outputs of deterministic systems driven by stochastic inputs. We will come back to this description a bit later, but first we will examine the scenario where a continuous-time stochastic signal is used as an input to a linear system, and describe the properties of the resulting output signal in terms of the properties of the input and of the system.

4.1 Linear time-invariant systems: a quick review

In ECE 210, we mostly deal with systems that are linear and time-invariant (LTI), because for such systems the relationship between the input and the output is particularly easy to write down. Each linear system is described by a function $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, such that, for any input signal x, the output is given by the *superposition integral*

$$\mathcal{S}[x](t) = \int_{-\infty}^{\infty} h(t,\tau)x(\tau)\mathrm{d}\tau \tag{4.1}$$

(as a memory refresher, prove that Eq. (4.1) specifies a linear system). The function h is called the *impulse response* of the system because it is explicitly given by

$$h(t,\tau) = \mathcal{S}[x](t) \qquad \text{when } x(t) = \delta(t-\tau).$$

When the system is also time-invariant, $h(t,\tau) = h(t+t_0, \tau+t_0)$ for any $t_0 \in \mathbb{R}$, so, overloading the notation a bit, we can rewrite (4.1) as a *convolution*:

$$\mathcal{S}[x](t) = h * x(t) = \int_{-\infty}^{\infty} h(t-\tau)x(\tau)\mathrm{d}\tau. \tag{4.2}$$

While computing convolutions is an important character-building part of ECE 210, things are much easier when we pass to the frequency domain using *Fourier transforms*. The Fourier transform of a function $g : \mathbb{R} \to \mathbb{R}$ (whenever it exists) is defined as

$$\widehat{g}(\omega) \triangleq \int_{-\infty}^{\infty} g(t)e^{-\mathrm{i}\omega t}\mathrm{d}t, \tag{4.3}$$

where the argument $\omega \in \mathbb{R}$ is the *frequency*[1] and $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. To get back to the time domain, we use the *Fourier inversion formula*

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{g}(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega. \tag{4.4}$$

The Fourier transform $\widehat{h}$ of the impulse response h is called the *transfer function* of the system, and the time-domain expression $y = h * x$ becomes

$$\widehat{y} = H\widehat{x}$$

in the frequency domain, where $\widehat{x}$ is the Fourier transform of the input x, and $\widehat{y}$ is the Fourier transform of the output y.

Remark on notation. In ECE 210, you were probably used to writing uppercase letters for Fourier transforms, like this: $X(\omega)$ instead of $\widehat{x}(\omega)$. However, this will cause confusion when we start

[1] Not to be confused with the generic element of the sample space Ω — hopefully, the meaning will be clear from the context.

dealing with stochastic inputs, since in that case X will already denote the time-domain input. I will still write H instead of $\widehat{h}$ because one must pay at least some respect to tradition, and because we will hardly ever use H to denote stochastic signals.

So, our next order of business is to adopt this formalism to the case when the input is a stochastic signal. But first we need to discuss the concept of stationarity.

4.2 Stationarity: weak and strong

As we have seen, both the Wiener process and the Poisson process have *stationary increments*. For example, if $W = (W_t)_{t\geq 0}$ is a Wiener process, then, for any two times $0 \leq s \leq t$ and for any $r \geq -s$, the distribution of the increment $W_t - W_s$ is the same as the distribution of the increment $W_{t+r} - W_{s+r}$, namely Gaussian with mean 0 and variance $D(t-s)$. Thus, the statistical properties of the increments of W are unaffected by time shifts.

This property turns out to be rather useful, so we abstract it into a definition: A stochastic signal $X = (X_t)_{t\in T}$ is *(strongly) stationary* if, for any $n \in \mathbb{N}$, any finite sequence of times $t_1, t_2, \ldots, t_n \in T$, and any $r \in \mathbb{R}$, such that $t_1 + r, \ldots, t_n + r \in T$,

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n}) \stackrel{\mathrm{d}}{=} (X_{t_1+r}, X_{t_2+r}, \ldots, X_{t_n+r}) \tag{4.5}$$

(the notation $U \stackrel{\mathrm{d}}{=} V$ means that the random quantities U and V have the same distribution). When X has a discrete state space X, the stationarity condition (4.5) means that

$$\mathbf{P}[X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n] = \mathbf{P}[X_{t_1+r} = x_1, X_{t_2+r} = x_2, \ldots, X_{t_n+r} = x_n]$$

for all $x_1, \ldots, x_n \in \mathsf{X}$; when X has a continuous state space X, (4.5) means that

$$\mathbf{P}[a_1 \leq X_{t_1} \leq b_1, \ldots, a_n \leq X_{t_n} \leq b_n] = \mathbf{P}[a_1 \leq X_{t_1+r} \leq b_1, \ldots, a_n \leq X_{t_n+r} \leq b_n]$$

for all subintervals $[a_1, b_1], \ldots, [a_n, b_n]$ of X. This definition covers discrete and continuous time. For example, any i.i.d. process $X = (X_k)_{k\in\mathbb{Z}_+}$ is strongly stationary.

However, for many purposes, strong stationarity is too much to ask for. Instead, we consider the following weaker notion: Let $X = (X_t)_{t\in T}$ be a stochastic signal with a continuous state space X. Then we say that X is *weakly stationary* (and write WS, for short) if it has the following two properties:

1. The mean function $m_X(t) = \mathbf{E}[X_t]$ is constant as a function of t:

$$m_X(t) = \mu, \qquad \forall t \in T. \tag{4.6}$$

2. For any two times $s, t \in T$ and any $r \in \mathbb{R}$ such that $s + r \in T$ and $t + r \in T$,

$$R_X(s,t) = R_X(s+r, t+r). \tag{4.7}$$

Here, $R_X(s,t) = \mathbf{E}[X_s X_t]$ is the autocorrelation function of X.

This property is much weaker than strong stationarity: one can easily construct examples of stochastic signals that are very nonstationary in the sense of (4.5), yet are weakly stationary. (Of course, any strongly stationary process with a continuous state space is also weakly stationary.) Now let us examine some implications of weak stationarity. While the definition applies to any T, we will focus on the easy case when T is closed under addition and subtraction: if $s,t \in T$, then $s+t \in T$ and $s-t \in T$. Then $0 \in T$. In that case, from (4.6) we get $\mu = \mathbb{E}[X_0]$, and from (4.7) with $r = -t$ we get

$$\mathbb{E}[X_t^2] = R_X(t,t) = R_X(0,0) = \mathbb{E}[X_0^2].$$

Thus, $\mathbb{E}[X_t^2] = \sigma^2$ for some $\sigma \geq 0$. Now, if we use (4.7) with $r = -s$ and $r = -t$, we have

$$R_X(s,t) = R_X(0,t-s) \qquad \text{and} \qquad R_X(t,s) = R_X(0,s-t).$$

Since $R_X(s,t) = R_X(t,s)$, we conclude that $R_X(s,t)$ depends only on $\tau = t-s$. Thus, for a weakly stationary stochastic process $X = (X_t)_{t\in T}$, we can overload the notation and write its autocorrelation function as $R_X(\tau)$. This really means that

$$R_X(\tau) = \mathbb{E}[X_t X_{t+\tau}] = \mathbb{E}[X_t X_{t-\tau}], \qquad \forall t, \tau \in T$$

and, in particular, implies that $R_X(\tau) = R_X(-\tau)$.

Before moving on, let us look at an example. Let A, B be two jointly distributed real-valued random variables, and consider the following stochastic signal $X = (X_t)_{t\in\mathbb{R}}$:

$$X_t = A\cos\omega t + B\sin\omega t \tag{4.8}$$

(here, ω is a deterministic angular frequency, not to be confused with a generic element of some probability space $(\Omega, \mathcal{F}, \mathbf{P})$). This is an example of a deterministic signal with stochastic parameters. What are the conditions on A and B for this signal to be weakly stationary? We claim that X is weakly stationary if and only if the following three conditions are satisfied:

1. $\mathbf{E}[A] = \mathbf{E}[B] = 0$ (both A and B have zero mean).
2. $\mathrm{Var}[A] = \mathrm{Var}[B] = \sigma^2$ (A and B have the same variance).
3. $\mathbf{E}[AB] = 0$ (A and B are uncorrelated).

We will only prove the statement that if X is WS, then A and B have to satisfy the above conditions; the converse will be a homework problem. So, suppose that X is WS. Then

$$m_X(t) = \mathbf{E}[X_t] = \mathbf{E}[A\cos\omega t + B\sin\omega t] = \mathbf{E}[A]\cos\omega t + \mathbf{E}[B]\sin\omega t,$$

and the only way for $m_X(t)$ to be a constant is to have $\mathbf{E}[A] = \mathbf{E}[B] = 0$, because otherwise it will depend on t. This proves Item 1. Now, again by the WS assumption, $\mathbf{E}[X_t^2] = R_X(0)$ must be a constant. Since $X_0 = A$ and $X_{\pi/2\omega} = B$, we have $\mathbf{E}[X_0^2] = \mathbf{E}[A^2] = \mathrm{Var}[A]$ and $\mathbf{E}[X_{\pi/2\omega}^2] = \mathbf{E}[B^2] =$

Var[B] (we have used the fact that both A and B have zero mean). Thus, Var[A] = Var[B] = σ^2 for some $\sigma \geq 0$. This proves Item 2. Finally, for any $t, \tau \in T$ we write

$$\begin{aligned} R_X(t, t+\tau) &= \mathbf{E}[X_t X_{t+\tau}] \\ &= \mathbf{E}\left[(A\cos\omega t + B\sin\omega t)\left(A\cos\omega(t+\tau) + B\sin\omega(t+\tau)\right)\right] \\ &= \mathbf{E}[A^2]\cos\omega t\cos\omega(t+\tau) + \mathbf{E}[B^2]\sin\omega t\sin\omega(t+\tau) \\ &\qquad + \mathbf{E}[AB]\left(\cos\omega t\sin\omega(t+\tau) + \sin\omega t\cos\omega(t+\tau)\right) \\ &= \sigma^2\cos\omega\tau + \mathbf{E}[AB]\sin\omega(2t+\tau), \end{aligned}$$

where we have used trigonometric identities and the fact that $\mathbf{E}[A^2] = \mathbf{E}[B^2] = \sigma^2$. We see that, unless $\mathbf{E}[AB] = 0$, $R_X(t, t+\tau)$ will depend on both t and τ, which would violate the assumption that X is WS. This proves Item 3.

As we will learn next, weak stationarity is preserved by linear time invariant (LTI) systems: if $X = (X_t)_{t\in R}$ is a WS input to an LTI system, then the output $Y = (Y_t)_{t\in R}$ is also WS, and its mean and correlation functions can be explicitly computed from those of X and from the impulse response of the system.

4.3 Systems with stochastic inputs: the LTI case

Consider an LTI system with impulse response h and fix a stochastic signal $X = (X_t)_{t\in\mathbb{R}}$. Then the output signal $Y = (Y_t)_{t\in\mathbb{R}}$ is related to X via the convolution integral, as in (4.2):

$$Y_t = \int_{-\infty}^{\infty} h(t-\tau)X_\tau \mathrm{d}\tau. \tag{4.9}$$

Of course, this is purely formal, since each X_τ is a random variable, and it takes some care to endow an integral like (4.9) with rirogous meaning. We will happily ignore all this fuss and just assume that the above integral is well-defined. Nevertheless, it is still a random quantity, and may not admit a closed-form expression. However, as we will now see, the output mean $m_Y(t)$, the input-output crosscorrelation $R_{XY}(s,t)$, and the output autocorrelation $R_Y(s,t)$ can be expressed in terms of the input mean $m_X(t)$, the input autocorrelation $R_X(s,t)$, and the impulse response h.

First, let's do the mean:

$$m_Y(t) = \mathbf{E}[Y_t] = \mathbf{E}\left[\int_{-\infty}^{\infty} h(t-\tau)X_\tau \mathrm{d}\tau\right]. \tag{4.10}$$

In general, interchaning expectations and integrals is a delicate matter, but we will be cavalier about it and just do it:

$$\mathbf{E}\left[\int_{-\infty}^{\infty} h(t-\tau)X_\tau \mathrm{d}\tau\right] = \int_{-\infty}^{\infty} \mathbf{E}[h(t-\tau)X_\tau]\mathrm{d}\tau = \int_{-\infty}^{\infty} h(t-\tau)\mathbf{E}[X_\tau]\mathrm{d}\tau.$$

Now, $\mathbf{E}[X_\tau] = m_X(\tau)$, so, putting everything together, we arrise at the pleasing formula

$$m_Y(t) = \int_{-\infty}^{\infty} h(t-\tau)m_X(\tau)\mathrm{d}\tau \equiv h * m_X(t). \tag{4.11}$$

That is, the output mean is given by the *convolution* of the input mean with the impulse response! We immediately note that if $m_X(t)$ is constant, i.e., $m_X(t) = \mu_X$ for all t, then $m_Y(t)$ is also constant and equal to

$$\mu_Y = \mu_X \int_{-\infty}^{\infty} h(t)\mathrm{d}t.$$

Here, we assume that the integral of h over the entire real line exists and is finite. We can also express this in terms of the transfer function H by noting that

$$\int_{-\infty}^{\infty} h(t)\mathrm{d}t = \int_{-\infty}^{\infty} h(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\Big|_{\omega=0} = H(0),$$

so we arrive at the following formula: if $m_X(t) = \mu_X$ for all t, then $m_Y(t) = H(0)\mu_X$.

Encouraged by our success, let's compute the input-output crosscorrelation next:

$$\begin{aligned} R_{XY}(s,t) &= \mathbf{E}[X_s Y_t] \\ &= \mathbf{E}\left[\int_{-\infty}^{\infty} h(t-\tau)X_s X_\tau \mathrm{d}\tau\right] \\ &= \int_{-\infty}^{\infty} h(t-\tau)\mathbf{E}[X_s X_\tau]\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty} h(t-\tau)R_X(s,\tau)\mathrm{d}\tau. \end{aligned}$$

This *almost* looks like a convolution, except that R_X is a function of two arguments. Note, however, that we are integrating over the *second* argument of R_X. So, we can write

$$R_{XY}(s,t) = h *_2 R_X(s,t),$$

where the subscript 2 on the asterisk indicates that we convolve only over the second argument of the input autocorrelation R_X:

$$h *_2 R_X(s,t) \triangleq \int_{-\infty}^{\infty} h(t-\tau)R_X(s,\tau)\mathrm{d}\tau.$$

This simplifies considerably when X is weakly stationary: in that case, $R_X(s,t) = R_X(t-s)$, and so

$$\begin{aligned} R_{XY}(s,t) &= h *_2 R_X(s,t) \\ &= \int_{-\infty}^{\infty} h(t-\tau)R_X(s,\tau)\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty} h(t-\tau)R_X(\tau-s)\mathrm{d}\tau \\ &= h * R_X(t-s) \end{aligned}$$

is just the ordinary convolution. Now we note that $R_{XY}(s,t)$ depends only on $t-s$, so we can write $R_{XY}(s,t) = R_{XY}(t-s)$. Consequently, we arrive at the formula

$$R_{XY}(\tau) = h * R_X(\tau), \tag{4.12}$$

where $R_{XY}(\tau) = \mathbb{E}[X_t Y_{t+\tau}]$ for all $t \in \mathbb{R}$. Thus, when the input X is weakly stationary, the input-output crosscorrelation is given by the convolution of the impulse response and the input autocorrelation!

Finally, let's look at the output autocorrelation. Again, blithely interchaning the order of expectation and integration, we have

$$\begin{aligned}
R_Y(s,t) &= \mathbf{E}[Y_s Y_t] \\
&= \mathbf{E}\left[\int_{-\infty}^{\infty} h(s-\tau)X_\tau d\tau \int_{-\infty}^{\infty} h(t-\tau')X_{\tau'}\mathrm{d}\tau'\right] \\
&= \mathbf{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s-\tau)h(t-\tau')X_\tau X_{\tau'}\mathrm{d}\tau\mathrm{d}\tau'\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s-\tau)h(t-\tau')\mathbf{E}[X_\tau X_{\tau'}]\mathrm{d}\tau\mathrm{d}\tau' \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s-\tau)h(t-\tau')R_X(\tau,\tau')\mathrm{d}\tau\mathrm{d}\tau'.
\end{aligned}$$

If we perform the integration over τ' first, we can recognize the $*_2$ operation:

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s-\tau)h(t-\tau')R_X(\tau,\tau')\mathrm{d}\tau\mathrm{d}\tau' &= \int_{-\infty}^{\infty} h(s-\tau)\underbrace{\left(\int_{-\infty}^{\infty} h(t-\tau')R_X(\tau,\tau')\mathrm{d}\tau'\right)}_{=h*_2R_X(\tau,t)}\mathrm{d}\tau \\
&= \int_{-\infty}^{\infty} h(s-\tau)R_{XY}(\tau,t)\mathrm{d}\tau.
\end{aligned}$$

So, if we now define the "partial convolution" $*_1$ over the first argument, we can write

$$R_Y(s,t) = h *_1 R_{XY}(s,t) = h *_1 (h *_2 R_X)(s,t).$$

This deceptively simple-looking formula hides a lot of complexity. But, once again, things simplify if X is weakly stationary. In that case, $h *_2 R_X(s,t) = h * R_X(t-s) = R_{XY}(t-s)$, and therefore

$$\begin{aligned}
h *_1 (h *_2 *R_X)(s,t) &= \int_{-\infty}^{\infty} h(s-\tau)R_{XY}(\tau,t)\mathrm{d}\tau \\
&= \int_{-\infty}^{\infty} h(s-\tau)R_{XY}(t-\tau)\mathrm{d}\tau.
\end{aligned} \tag{4.13}$$

Again, this integral looks very much like a convolution, but here we hit a snag. In general, a convolution integral of two functions f and g will look like this:

$$\int_{-\infty}^{\infty} f(t-\tau)g(\tau-s)\mathrm{d}\tau,$$

i.e., if we add the arguments of f and g in the integrand, the integration variable τ will cancel, and it is not hard to show that the integral is equal to $f * g(t-s)$. But if we add up the arguments of h and R_{XY} in the integrand of (4.13), we get $(s-\tau)+(t-\tau) = s+t-2\tau$, and τ most certainly does not cancel! Fortunately, there is a nice hack out of this conundrum: given h, define another function $\tilde{h}$ by setting $\tilde{h}(t) \triangleq h(-t)$. For obvious reasons, we call $\tilde{h}$ the *time reversal* of h. Then, with this definition, we can rewrite (4.13) as

$$\int_{-\infty}^{\infty} h(s-\tau)R_{XY}(t-\tau)\mathrm{d}\tau = \int_{-\infty}^{\infty} \tilde{h}(\tau-s)R_{XY}(t-\tau)\mathrm{d}\tau,$$

and behold: $(\tau-s)+(t-\tau) = t-s$, and so we can write $h *_1 (h *_2 R_X)(s,t) = \tilde{h} * h * R_X(t-s)$. Thus, when the input X is weakly stationary, the output autocorrelation is given by the double convolution:

$$R_Y(\tau) = \tilde{h} * h * R_X(\tau), \tag{4.14}$$

i.e., if the input X is weakly stationary with autocorrelation $R_X(\tau)$, then the output Y is also weakly stationary. Moreover, in that case the crosscorrelation $R_{XY}(s,t)$ also depends only on $t-s$, so we say that the input and the output are *jointly weakly stationary*: each of them is WS, and $R_{XY}(s,t) = R_{XY}(t-s)$ [although be careful: $R_{XY}(t-s) = \mathbf{E}[X_s Y_t] \neq \mathbf{E}[X_t Y_s] = R_{XY}(s-t)$].

4.4 Input-output relations in the frequency domain: power spectral densities

No one likes to compute convolution integrals (even though it builds character), so we pass to the frequency domain. Let X be a WS stochastic signal with autocorrelation $R_X(\tau)$. The *power spectral density* of X, denoted by S_X, is simply the Fourier transform of R_X:

$$S_X(\omega) \triangleq \widehat{R_X}(\omega) = \int_{-\infty}^{\infty} R_X(\tau)e^{-\mathrm{i}\omega\tau}\mathrm{d}\tau. \tag{4.15}$$

The autocorrelation function can be recovered from the power spectral density via the Fourier inversion formula (4.4):

$$R_X(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)e^{\mathrm{i}\omega\tau}\mathrm{d}\omega. \tag{4.16}$$

The word "spectral" evokes frequency content, so this makes sense. But what do "power" and "density" mean? If we think about X_t as a (random) current passing through a unit resistance, then the voltage across the resistor is also equal to X_t, by Ohm's law. Consequently, the dissipated power at time t is equal to

$$(\text{voltage at time } t)\cdot(\text{current at time } t) = X_t^2.$$

Thus, we may think of $\mathbf{E}[X_t^2]$ as the *average power* dissipated at time t. Now, if X is WS, then $\mathbf{E}[X_t^2] = \mathbf{E}[X_t X_t] = R_X(0)$. On the other hand, if we substitute $\tau = 0$ into the Fourier inversion formula (4.16), we get

$$\mathbf{E}[X_t^2] = R_X(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)\mathrm{d}\omega. \tag{4.17}$$

Similarly, given two jointly WS stochastic signals X and Y, we define their *cross-power spectral density* S_{XY} as the Fourier transform of the crosscorrelation function R_{XY}:

$$S_{XY}(\omega) = \widehat{R_{XY}}(\omega) = \int_{-\infty}^{\infty} R_{XY}(\tau) e^{\mathrm{i}\omega\tau} \mathrm{d}\tau.$$

Using power spectral densities, we can obtain frequency-domain forms of the results of the preceding section. For example, taking the Fourier transform of both sides of (4.12), we get

$$S_{XY}(\omega) = H(\omega) S_X(\omega). \tag{4.18}$$

Similarly, taking the Fourier transform of both sides of (4.14), we get

$$S_Y(\omega) = \tilde{H}(\omega) H(\omega) S_X(\omega), \tag{4.19}$$

where $\tilde{H}$ is the Fourier transform of the time-reversed impulse response $\tilde{h}$. However, we can express $\tilde{H}$ in terms of H:

$$\begin{aligned}
\tilde{H}(\omega) &= \int_{-\infty}^{\infty} \tilde{h}(\tau) e^{-\mathrm{i}\omega\tau} \mathrm{d}\tau \\
&= \int_{-\infty}^{\infty} h(-\tau) e^{-\mathrm{i}\omega\tau} \mathrm{d}\tau \\
&= \int_{-\infty}^{\infty} h(\tau) e^{\mathrm{i}\omega\tau} \mathrm{d}\tau \\
&= H(-\omega),
\end{aligned}$$

where in the penultimate line we have made the change of variable $\tau \to -\tau$. Now, since h is real-valued, we see that $\tilde{H}(\omega) = H(-\omega)$ is just the complex conjugate of $H(\omega)$: $\tilde{H}(\omega) = H^*(\omega)$. Therefore, (4.19) simplifies to

$$S_Y(\omega) = |H(\omega)|^2 S_X(\omega). \tag{4.20}$$

Thus, when a WS stochastic signal is used as an input to an LTI system, the effect in the frequency domain is to reshape the power spectral density in proportion to $|H|^2$.

Example: white and colored noise. Consider a zero-mean WS stochastic signal Z with the flat power spectral density $S_Z(\omega) = q$, where $q > 0$ is some fixed constant. By analogy with white visible light that contains all spectral components, we call such a signal *white noise* with intensity q. The autocorrelation function is then given by $R_Z(\tau) = q\delta(\tau)$. Since Z is zero-mean,

$$R_Z(\tau) = C_Z(\tau) = \mathbf{E}[Z_t Z_{t+\tau}]$$

for any $t, \tau \in \mathbb{R}$. In this case, $R_Z(\tau) = 0$ for all $\tau > 0$, i.e., Z_t and $Z_{t'}$ are uncorrelated for all $t \neq t'$. Although white noise is a useful mathematical construction (for example, many stochastic signals

of interest can be interpreted as outputs of deterministic systems driven by white-noise inputs), it is unphysical. To see why, let us substitute $S_Z(\omega) = q$ into (4.17) to get

$$\mathbf{E}[Z_t^2] = \frac{1}{2\pi}\int_{-\infty}^{\infty} q \, \mathrm{d}\omega = +\infty.$$

That is, pure white noise has infinite average power! Of course, we can simply continue using it while remaining mindful of its unphysical nature (just like we do with the unit impulse), or we can develop various approximations. For example, it is often possible to consider stochastic signals whose power spectra are flat and nonzero only in some band of frequencies. Thus, consider a WS stochastic signal Z with

$$S_Z(\omega) = \begin{cases} q, & -\omega_0 \le \omega \le \omega_0 \\ 0, & \text{otherwise} \end{cases}, \tag{4.21}$$

where $q > 0$ is the intensity parameter and $\omega_0 > 0$ is some cutoff frequency. Then

$$\mathbf{E}[Z_t^2] = \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0} q \, \mathrm{d}\omega = \frac{q\omega_0}{\pi},$$

which is finite. We can also compute the autocorrelation function of Z:

$$\begin{aligned} R_Z(\tau) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_Z(\omega) e^{\mathrm{i}\omega\tau} \mathrm{d}\omega \\ &= \frac{q}{2\pi}\int_{-\omega_0}^{\omega_0} e^{\mathrm{i}\omega\tau} \mathrm{d}\omega \\ &= \frac{q}{2\pi} \cdot \frac{1}{\mathrm{i}\tau} e^{\mathrm{i}\omega\tau}\Big|_{-\omega_0}^{\omega_0} \\ &= \frac{q}{\pi\tau} \cdot \frac{e^{\mathrm{i}\omega_0\tau} - e^{-\mathrm{i}\omega_0\tau}}{2\mathrm{i}} \\ &= \frac{q \sin\omega_0\tau}{\pi\tau}. \end{aligned}$$

Recalling the definition of the sinc function[2]

$$\operatorname{sinc}(u) = \begin{cases} \frac{\sin \pi u}{\pi u}, & u \neq 0 \\ 1, & u = 0 \end{cases}$$

we can write

$$R_Z(\tau) = 2qf_0 \operatorname{sinc}(2f_0\tau), \tag{4.22}$$

where $f_0 = \omega_0/2\pi$ is the cutoff frequency in Hz. A plot of R_Z with $q = 5$ and $\omega_0 = 4$ rad/s is shown in Fig. 1. Note that because $|R_Z(\tau)| > 0$ for all τ, there are correlations between Z_t and $Z_{t+\tau}$, but they decay to zero as $|\tau| \to \infty$.

[2]This is the engineer's sinc function. The mathematician's sinc function is $\operatorname{sinc}(u) = \frac{\sin u}{u}$.

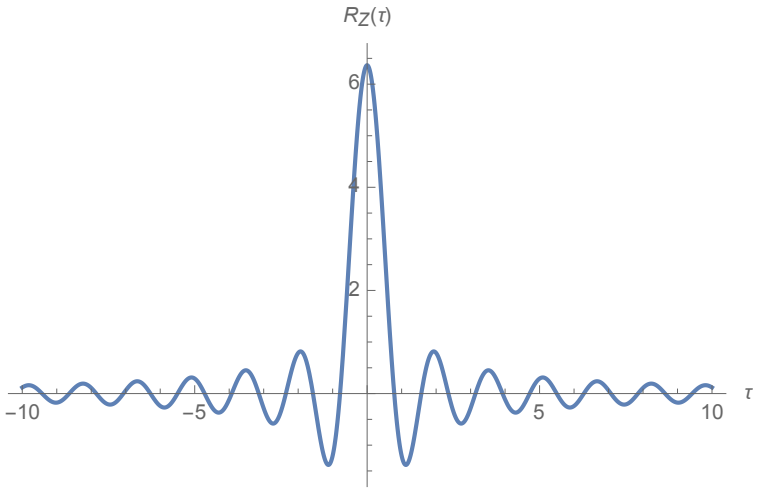

Figure 1: The sinc autocorrelation function.

The term "colored noise" refers to stochastic signals whose power spectra are roughly shaped like the corresponding colors of the visible spectrum. For example, the term "pink noise" refers to power spectra of the form $S(\omega) \propto 1/\omega$. Later on, we will return to this noise model when discussing flicker noise in electronic devices.

4.4.1 Properties of power spectral densities

The power spectral density S_X of any WS stochastic signal $X = (X_t)_{t\in\mathbb{R}}$ is real (i.e., $S_X(\omega) \in \mathbb{R}$ for all ω), even (i.e., $S_X(-\omega) = S_X(\omega)$ for all ω), and nonnegative (i.e., $S_X(\omega) \geq 0$ for all ω).

Indeed, recall that S_X is the Fourier transform of the autocorrelation function R_X. The autocorrelation function R_X takes real values, and it is also even:

$$R_X(-\tau) = \mathbb{E}[X_t X_{t-\tau}] = \mathbb{E}[X_{t+\tau} X_t] = R_X(\tau).$$

Therefore, using Euler's formula, we have

$$\begin{aligned} S_X(\omega) &= \int_{-\infty}^{\infty} R_X(\tau) e^{-\mathrm{i}\omega\tau} \mathrm{d}\tau \\ &= \int_{-\infty}^{\infty} R_X(\tau) \cos\omega\tau \mathrm{d}\tau + \mathrm{i} \int_{-\infty}^{\infty} R_X(\tau) \sin\omega\tau \mathrm{d}\tau. \end{aligned}$$

Since $R_X(\tau)$ is even and $\sin\omega\tau$ is odd, the second integral is identically zero, and we obtain the formula

$$S_X(\omega) = \int_{-\infty}^{\infty} R_X(\tau) \cos\omega\tau \mathrm{d}\tau. \tag{4.23}$$

Since the integrand takes real values, $S_X(\omega) \in \mathbb{R}$ for all ω. Moreover, the integrand is even as a function of ω, so S_X is also an even function. This also implies that R_X is given by

$$\begin{aligned} R_X(\tau) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega) e^{\mathrm{i}\omega\tau}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)\cos\omega\tau\mathrm{d}\omega + \frac{\mathrm{i}}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)\sin\omega\tau\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)\cos\omega\tau\mathrm{d}\omega, \end{aligned}$$

where we have used the same reasoning as before: $S_X(\omega)\sin\omega\tau$ is an odd function of ω, so the second integral is zero.

It remains to show that $S_X(\omega) \geq 0$ for all ω. Let X be the input to a bandpass filter with the transfer function

$$H(\omega) = \begin{cases} 1, & \omega_0 \leq |\omega| \leq \omega_1 \\ 0, & \text{otherwise} \end{cases},$$

where $\omega_0 \leq \omega_1$ are arbitrary nonnegative constants. Let Y denote the output stochastic signal. Then

$$S_Y(\omega) = |H(\omega)|^2 S_X(\omega) = \begin{cases} S_X(\omega), & \omega_0 \leq |\omega| \leq \omega_1 \\ 0, & \text{otherwise} \end{cases}.$$

Now, using the Fourier inversion formula and the evenness of S_X,

$$\begin{aligned} R_Y(0) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_Y(\omega)\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\omega_0}^{-\omega_1} S_X(\omega)\mathrm{d}\omega + \frac{1}{2\pi}\int_{\omega_0}^{\omega_1} S_X(\omega)\mathrm{d}\omega \\ &= \frac{1}{\pi}\int_{\omega_0}^{\omega_1} S_X(\omega)\mathrm{d}\omega. \end{aligned}$$

Since $R_Y(0) = \mathbf{E}[Y_t^2] \geq 0$, we see that

$$\int_{\omega_0}^{\omega_1} S_X(\omega)d\omega \geq 0, \qquad \forall \omega_0, \omega_1. \tag{4.24}$$

The only way for (4.24) to hold is if $S_X(\omega) \geq 0$ for all ω.

In fact, for any function S of frequency ω which is real, even, nonnegative, and satisfies the condition

$$a \triangleq \int_{-\infty}^{\infty} S(\omega)\mathrm{d}\omega < \infty,$$

we can construct a WS stochastic signal X, such that $S = S_X$. To prove this, consider the following stochastic signal:

$$X_t = \cos(\Omega t + \Theta),$$

where Ω and Θ are two mutually independent random variables. Let us assume that $a = \pi$ (otherwise, we can simply rescale S). We will take $\Theta \sim \text{Uniform}(0, 2\pi)$, and we will choose the pdf f_Ω of Ω later. First, let us prove that X is WS. For that, we have to compute its mean and autocorrelation functions. Since Ω and Θ are independent, we can write

$$\begin{aligned} m_X(t) &= \mathbf{E}[X_t] \\ &= \mathbf{E}[\cos(\Omega t + \Theta)] \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_0^{2\pi} f_\Omega(\omega)\cos(\omega t + \theta)\,\mathrm{d}\theta\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f_\Omega(\omega)\left(\int_0^{2\pi}\cos(\omega t + \theta)\mathrm{d}\theta\right)\mathrm{d}\omega. \end{aligned}$$

The integral in parentheses is equal to zero, so $m_X(t) = 0$ for all t. For the autocorrelation function, using the trigonometric identity $\cos u \cos v = \frac{1}{2}[\cos(u+v) + \cos(u-v)]$, we have

$$\begin{aligned} R_X(t, t+\tau) &= \mathbf{E}[X_t X_{t+\tau}] \\ &= \mathbf{E}[\cos(\Omega t + \Theta)\cos(\Omega(t+\tau) + \Theta)] \\ &= \frac{1}{2}\mathbf{E}\left[\cos(2\Omega t + \Omega\tau + \Theta) + \cos(\Omega\tau)\right] \\ &= \frac{1}{2}\mathbf{E}[\cos(2\Omega t + \Omega\tau + \Theta)] + \frac{1}{2}\mathbf{E}[\cos\Omega\tau]. \end{aligned}$$

Using the same reasoning as in the derivation of m_X, we see that the first expectation is identically zero, and therefore

$$R_X(t, t+\tau) = \frac{1}{2}\mathbf{E}[\cos\Omega\tau]. \tag{4.25}$$

Since $m_X = 0$ and $R_X(t, t+\tau)$ depends only on τ, X is indeed WS. Now we will pick f_Ω to guarantee that $S_X = S$. To that end, we will use the Fourier inversion formula

$$R_X(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)\cos(\omega\tau)\mathrm{d}\omega$$

(recall that S_X is even). On the other hand, from (4.25) we have

$$R_X(\tau) = \frac{1}{2}\int_{-\infty}^{\infty} f_\Omega(\omega)\cos(\omega\tau)\mathrm{d}\omega.$$

In particular,

$$R_X(0) = \frac{1}{2} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)\mathrm{d}\omega$$

Let us take $f_\Omega(\omega) = \frac{S(\omega)}{\pi}$. This function is nonnegative and integrates to 1. Therefore, it is a valid pdf. By uniqueness of Fourier transforms, we have ensured that $S = \widehat{R_X}$, and therefore $S_X = S$.

4.5 Gaussian stochastic signals and Bussgang's theorem

An important class of stochastic signals consists of *Gaussian stochastic signals* (or Gaussian processes). To define a Gaussian stochastic signal, we first recall the definition of a Gaussian random variable: A random variable X is called Gaussian if its characteristic function $\Phi_X(u) \triangleq \mathbf{E}[e^{\mathrm{i}uX}]$ has the form

$$\Phi_X(u) = \exp\left(\mathrm{i}\mu u - \frac{u^2\sigma^2}{2}\right) \tag{4.26}$$

for some $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$. It can be shown then that $\mu = \mathbf{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$ (in the degenerate case $\sigma = 0$, $X = \mu$ with probability one). When $\sigma > 0$, the pdf of X is obtained by taking the inverse Fourier transform of Φ_X:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_X(u) e^{-\mathrm{i}ux}\mathrm{d}u \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \end{aligned}$$

When $\sigma = 0$, X does not have a pdf.

Now consider a *random vector*, i.e., a stochastic signal of the form $X = (X_t)_{t\in T}$ with $T = \{1,\ldots,n\}$. We can represent such signals by column vectors, as in

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

In the same way, we can represent the mean and the autocovariance functions of X:

$$m_X = \begin{pmatrix} m_X(1) \\ m_X(2) \\ \vdots \\ m_X(n) \end{pmatrix} \quad \text{and} \quad C_X = \begin{pmatrix} C_X(1,1) & C_X(1,2) & \ldots & C_X(1,n) \\ C_X(2,1) & C_X(2,2) & \ldots & C_X(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ C_X(n,1) & C_X(n,2) & \ldots & C_X(n,n) \end{pmatrix},$$

where, as usual, $M_X(t) = \mathbf{E}[X_t]$ and $C_X(s,t) = \mathbf{E}[X_sX_t] - \mathbf{E}[X_s]\mathbf{E}[X_t]$. The $n \times n$ *covariance matrix* C_X is symmetric, since $C_X(s,t) = C_X(t,s)$, and the diagonal entries are equal to the variances of the X_t's:

$$C_X(t,t) = \mathbb{E}[X_t^2] - (\mathbb{E}[X_t])^2 = \mathrm{Var}[X_t].$$

Now let $u = (u_1,\ldots,u_n)^T$ be a fixed deterministic vector, and consider the scalar random variable $Z = u^TX = u_1X_1 + \ldots + u_nX_n$. Geometrically, Z is the *projection* of the random vector X onto the deterministic vector u. We can easily compute its mean and variance:

$$\mathbf{E}[Z] = \mathbf{E}\left[\sum_{t=1}^{n} u_tX_t\right] = \sum_{t=1}^{n} u_t\mathbf{E}[X_t] = \sum_{t=1}^{n} u_tm_X(t) = u^Tm_X$$

and

$$\begin{aligned}\mathrm{Var}[Z] &= \mathbf{E}[Z^2] - (\mathbf{E}Z)^2 \\ &= \sum_{s=1}^{n}\sum_{t=1}^{n} (\mathbf{E}[u_s u_t X_s X_t] - u_s u_t \mathbf{E}[X_s]\mathbf{E}[X_t]) \\ &= \sum_{s=1}^{n}\sum_{t=1}^{n} u_s u_t (\mathbf{E}[X_s X_t] - \mathbf{E}[X_s]\mathbf{E}[X_t]) \\ &= \sum_{s=1}^{n}\sum_{t=1}^{n} u_s u_t C_X(s,t) \\ &= u^T C_X u.\end{aligned}$$

We now introduce the following definition: we say that X is a *Gaussian random vector* (or that $X_1, \ldots, X_n$ are *jointly Gaussian random variables*) if the scalar random variable $Z = u^T X$ is Gaussian for each choice of the vector of coefficients $u \in \mathbb{R}^n$. In light of our definition of a Gaussian random variable, this is equivalent to the following: X is a Gaussian random vector if, for any $u \in \mathbb{R}^n$, the characteristic function $\Phi_Z = \Phi_{u^T X}$ of $Z = u^T X$ takes the form

$$\begin{aligned}\Phi_Z(\alpha) &= \exp\left(\mathrm{i}\alpha \mathbf{E}[Z] - \frac{\alpha^2 \mathrm{Var}[Z]}{2}\right) \\ &= \exp\left(\mathrm{i}\alpha u^T m_X - \frac{\alpha^2 u^T C_X u}{2}\right)\end{aligned}$$

for all $\alpha \in \mathbb{R}$. Note that

$$\Phi_{u^T X}(\alpha) = \mathbf{E}[e^{\mathrm{i}\alpha u^T X}] = \mathbf{E}[e^{\mathrm{i}(\alpha u)^T X}] = \Phi_{\alpha u^T X}(1),$$

so specifying the characteristic functions of all linear combinations $u^T X$ is equivalent to specifying the *joint characteristic function* of X, which is a function of a vector argument $u \in \mathbb{R}^n$ and takes the form

$$\Phi_X(u) \equiv \Phi_{X_1,\ldots,X_n}(u_1, \ldots, u_n) \triangleq \mathbf{E}[e^{\mathrm{i}u^T X}]. \tag{4.27}$$

Just like the characteristic function of a scalar random variable X uniquely specifies the distribution of X, the joint characteristic function of a random vector $X = (X_1, \ldots X_n)^T$ uniquely specifies the joint distribution of its coordinates $X_1, \ldots, X_n$. In particular, if the X_t's have a joint pdf, then it is given by the multidimensional Fourier inversion formula

$$f_X(x) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Phi_X(u) e^{-\mathrm{i}u^T X} \mathrm{d}u_1 \ldots \mathrm{d}u_n. \tag{4.28}$$

Because $\Phi_X(u) = \Phi_{u^T X}(1)$, we see that if we know the probability distributions of all of its projections $Z = u^T X$, $u \in \mathbb{R}^n$, then we know the joint probability distribution of $X_1, \ldots, X_n$ (this is referred to

as the *Cramèr–Wold theorem*). Consequently, a random vector X is Gaussian if and only if its joint characteristic function takes the form

$$\Phi_X(u) = \exp\left(\mathrm{i}u^T m_X - \frac{1}{2}u^T C_X u\right), \qquad u \in \mathbb{R}^n. \tag{4.29}$$

Along the way, we have established a nice fact: the projection of a Gaussian random vector onto a deterministic vector is always Gaussian.

Important! You should always keep in mind the crucial distinction between a *Gaussian random vector* and a vector of Gaussian random variables. It is easy to construct examples of random vectors $X = (X_1, \ldots, X_n)^T$, where each X_t is Gaussian, yet $X_1, \ldots, X_n$ are not jointly Gaussian (you will see such an example in the homework). If $X_1, \ldots, X_n$ are not jointly Gaussian, then the projection $u^T X$ may not be Gaussian.

We are now ready to define a general Gaussian stochastic signal: a stochastic signal $X = (X_t)_{t\in T}$ is Gaussian if, for any $n \in \mathbb{N}$ and any $t_1, \ldots, t_n \in T$, the random variables $X_{t_1}, \ldots, X_{t_n}$ are jointly Gaussian. That is, their joint characteristic function has the form

$$\Phi_{X_{t_1},\ldots,X_{t_n}}(u_1, \ldots, u_n) = \exp\left(\mathrm{i}\sum_{k=1}^n u_k m_X(t_k) - \frac{1}{2}\sum_{j=1}^n\sum_{k=1}^n u_j u_k C_X(t_j, t_k)\right). \tag{4.30}$$

This implies that a weakly stationary Gaussian stochastic signal is, in fact, strongly stationary. Indeed, suppose that X is WS. Then $\mu_X(t) = \mu$ for all t, and $C_X(s,t) = C_X(t-s)$ for all s, t. Using these facts in (4.30), we obtain

$$\Phi_{X_{t_1},\ldots,X_{t_n}}(u_1, \ldots, u_n) = \exp\left(\mathrm{i}\mu\sum_{k=1}^n u_k - \frac{1}{2}\sum_{j=1}^n\sum_{k=1}^n u_j u_k C_X(t_k - t_j)\right). \tag{4.31}$$

Now fix an arbitrary $\tau \in \mathbb{R}$ and consider the ranodm variables $X_{t_1+\tau}, \ldots, X_{t_n+\tau}$. Since $\mu_X(t)$ is constant, and since $C_X(s+\tau, t+\tau) = C_X(t-s)$ their joint characteristic function is also given by (4.31):

$$\Phi_{X_{t_1+\tau},\ldots,X_{t_n+\tau}}(u_1, \ldots, u_n) = \exp\left(\mathrm{i}\mu\sum_{k=1}^n u_k - \frac{1}{2}\sum_{j=1}^n\sum_{k=1}^n u_j u_k C_X(t_k - t_j)\right). \tag{4.32}$$

Thus, for any choice of $t_1, \ldots, t_n$ and τ, $(X_{t_1}, \ldots, X_{t_n})$ and $(X_{t_1+\tau}, \ldots, X_{t_n+\tau})$ have the same joint characteristic function and hence the same distribution. Therefore X is strongly stationary. For this reason, we can just say "stationary" when dealing with Gaussian stochastic signals.

Now that we have collected some background about Gaussian stochastic signals, we can state and prove a remarkable result called *Bussgang's theorem*.[3] Let X be a stationary Gaussian stochastic

[3] J.J. Bussgang, "Cross-correlation function of amplitude-distorted Gaussian signals," MIT Research Laboratory of Electronics, Technical Report 216, March 1952.

signal with zero mean, and let Y be obtained by passing X through a memoryless nonlinearity. That is, for each $t \in \mathbb{R}$, we have $Y_t = g(X_t)$, where $g : \mathbb{R} \to \mathbb{R}$ is a fixed transformation. Suppose that g is differentiable[4] and has moderate growth at infinity, i.e.,

$$\lim_{|x|\to\infty} g(x)e^{-ax^2} = 0, \qquad \forall a > 0. \tag{4.33}$$

For example, any polynomial in x would do. Bussgang's theorem gives the following expression for the crosscorrelation function between X and Y:

$$R_{XY}(t, t+\tau) = \mathbf{E}[X_t Y_{t+\tau}] = K \cdot R_X(\tau), \tag{4.34}$$

where

$$K = \mathbf{E}[g'(X_0)].$$

Since $R_{XY}(t, t+\tau)$ depends only on τ, we can write $R_{XY}(\tau) = KR_X(\tau)$ instead. The remarkable thing about (4.34) is its simplicity, despite the fact that g may be highly nonlinear and the output stochastic signal Y may not even be Gaussian. Before giving the proof of (4.34), let us take a look at a couple of examples.

Hard limiter. Consider the following nonlinear transformation:

$$g(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}.$$

We can write g as a difference of two unit-step functions: $g(x) = u(x) - u(-x)$. Therefore, $g'(x) = 2\delta(x)$. Let X be a stationary Gaussian stochastic signal, and let f_0 denote the pdf of X_0. Then

$$f_0(x) = \frac{1}{\sqrt{2\pi R_X(0)}} e^{-x^2/2R_X(0)},$$

and therefore

$$\begin{aligned} \mathbf{E}[g'(X_0)] &= 2\int_{-\infty}^{\infty} f_0(x)\delta(x)\mathrm{d}x \\ &= 2f_0(0) \\ &= \sqrt{\frac{2}{\pi R_X(0)}}. \end{aligned}$$

Hence, Bussgang's theorem gives

$$R_{XY}(\tau) = \sqrt{\frac{2}{\pi R_X(0)}} R_X(\tau).$$

[4] We understand differentiability in the generalized sense and allow functions like the unit step or the unit ramp.

Half-wave rectifier. Consider

$$g(x) = \max\{x, 0\} = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}.$$

Then $g'(x) = u(x)$, so for any zero-mean Gaussian random variable X we have

$$\mathbf{E}[g'(X)] = \mathbf{E}[u(X)] = \int_0^\infty f_X(x)\mathrm{d}x = \frac{1}{2}.$$

Consequently, Bussgang's theorem gives $R_{XY}(\tau) = \frac{1}{2}R_X(\tau)$.

Now we give the proof of Bussgang's theorem. First, we note that it is equivalent to the following statement: Let U and V be two jointly Gaussian random variables with zero mean. Then for any g satisfying (4.33) we have

$$\mathbf{E}[Ug(V)] = \mathrm{Cov}(U, V)\mathbf{E}[g'(V)]. \tag{4.35}$$

We obtain (4.34) by applying (4.35) to $U = X_t$ and $V = X_{t+\tau}$. Since X is zero-mean, $\mathrm{Cov}(X_t, X_{t+\tau}) = C_X(\tau) = R_X(\tau)$, and, since X is stationary, $\mathbf{E}[g'(X_{t+\tau})] = \mathbf{E}[g'(X_0)]$. Thus, we proceed to prove (4.35). We will derive it from another important result, known as *Stein's identity*: Let g be a differentiable function satisfying (4.33), and let U be a zero-mean Gaussian random variable with variance σ^2. Then

$$\mathbf{E}[Ug(U)] = \sigma^2\,\mathbf{E}[g'(U)]. \tag{4.36}$$

Remark on Stein's identity. The importance of Stein's identity (4.36) lies in the fact that it actually fully characterizes the Gaussian distribution: It can be proved that any zero-mean, unit-variance random variable U, such that

$$\mathbf{E}[Ug(U)] = \mathbf{E}[g'(U)]$$

holds for all smooth g, is Gaussian. This can be pushed further to imply that any random variable U with zero mean and unit variance, such that

$$\mathbf{E}[Ug(U) - g'(U)] \approx 0$$

for all smooth g, is "close to Gaussian."

To prove (4.36), we note the following about the pdf of U:

$$f_U'(u) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{\mathrm{d}}{\mathrm{d}u}e^{-u^2/2\sigma^2} = -\frac{u}{\sigma^2}f_U(u).$$

Therefore,

$$\begin{aligned}\mathbf{E}[Ug(U)] &= \int_{-\infty}^\infty ug(u)f_U(u)\mathrm{d}u \\ &= -\sigma^2\int_{-\infty}^\infty g(u)f_U'(u)\mathrm{d}u.\end{aligned}$$

Integrating by parts, we get

$$\mathbf{E}[Ug(U)] = -\sigma^2 g(u) f_U(u)\Big|_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} g'(u) f_U(u) \mathrm{d}u.$$

Since g satisfies (4.33), the first term on the right-hand side is identically zero, and we obtain

$$\mathbf{E}[Ug(U)] = \sigma^2 \int_{-\infty}^{\infty} g'(u) f_U(u) \mathrm{d}u = \sigma^2 \mathbf{E}[g'(U)].$$

Now we will use Stein's identity to prove (4.35). To that end, we first use a trick which you can think of as "Gram-Schmidt orthogonalization for Gaussians:" Given the pair (U, V), define the random variable

$$\tilde{V} \triangleq V - \frac{c}{\sigma_U^2} U,$$

where $c = \mathrm{Cov}(U, V) = \mathbf{E}[UV]$ and $\sigma_U^2 = \mathrm{Var}[U]$. Since U and V are jointly Gaussian with zero mean, $\tilde{V}$ is also Gaussian with zero mean. Moreover, U and $\tilde{V}$ are *uncorrelated*:

$$\begin{aligned}
\mathbf{E}[U\tilde{V}] &= \mathbf{E}\left[U\left(V - \frac{c}{\sigma_U^2} U\right)\right] \\
&= \mathbf{E}[UV] - \frac{c}{\sigma_U^2} \mathbf{E}[U^2] \\
&= \mathbf{E}[UV] - \frac{c}{\sigma_U^2} \sigma_U^2 \\
&= 0.
\end{aligned}$$

As you will prove in the homework, if two jointly Gaussian random variables are uncorrelated, then they are also independent. Therefore, U and $\tilde{V}$ are independent. Using this fact, we can write

$$\begin{aligned}
\mathbf{E}[Ug(V)] &= \mathbf{E}\left[Ug\left(\tilde{V} + \frac{c}{\sigma_U^2} U\right)\right] \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u g\left(\tilde{v} + \frac{c}{\sigma_U^2} u\right) f_U(u) f_{\tilde{V}}(\tilde{v}) \mathrm{d}u \, \mathrm{d}\tilde{v}.
\end{aligned}$$

We will first integrate over u, and then over $\tilde{v}$:

$$\begin{aligned}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u g\left(\tilde{v} + \frac{c}{\sigma_U^2} u\right) f_U(u) f_{\tilde{V}}(\tilde{v}) \mathrm{d}u \, \mathrm{d}\tilde{v} &= \int_{-\infty}^{\infty} f_{\tilde{V}}(\tilde{v}) \left[\int_{-\infty}^{\infty} u g\left(\tilde{v} + \frac{c}{\sigma_U^2} u\right) f_U(u) \mathrm{d}u\right] \mathrm{d}\tilde{v} \\
&= \int_{-\infty}^{\infty} f_{\tilde{V}}(\tilde{v}) \mathbf{E}\left[Ug\left(\tilde{v} + \frac{c}{\sigma_U^2} U\right)\right] \mathrm{d}\tilde{v}. \qquad (4.37)
\end{aligned}$$

Using Stein's identity and the chain rule for differentiation, we can rewrite the expectation in (4.37) as follows:

$$\mathbf{E}\left[Ug\left(\tilde{v} + \frac{c}{\sigma_U^2} U\right)\right] = \sigma_U^2 \cdot \frac{c}{\sigma_U^2} \mathbf{E}\left[g'\left(\tilde{v} + \frac{c}{\sigma_U^2} U\right)\right] = c\mathbf{E}\left[g'\left(\tilde{v} + \frac{c}{\sigma_U^2} U\right)\right].$$

Substituting this expression back into (4.37), we get

$$\begin{aligned}\mathbf{E}[Ug(V)] &= c\int_{-\infty}^{\infty} f_{\tilde{V}}(\tilde{v})\mathbf{E}\left[g'\left(\tilde{v}+\frac{c}{\sigma_U^2}U\right)\right]\mathrm{d}\tilde{v}\\ &= c\mathbf{E}\left[g'\left(\tilde{V}+\frac{c}{\sigma_U^2}U\right)\right]\\ &= c\mathbf{E}[g'(V)]\\ &= \mathrm{Cov}(U,V)\cdot\mathbf{E}[g'(V)],\end{aligned}$$

which proves (4.35) and thus Bussgang's theorem.

4.6 Poisson point processes and Campbell's theorem

Gaussian processes provide a basic model of stochastic phenomena involving continuously fluctuating quantities. Poisson point processes, on the other hand, serve as a good model of phenomena involving discrete events occurring at random times. Recall that a Poisson point process with rate λ is a discrete-time stochastic signal $(T_k)_{k\in\mathbb{N}}$, where the random arrival (or event occurrence) times $0 < T_1 < T_2 < T_3 < \ldots$ are such that the interarrival times $Z_k = T_k - T_{k-1}$ (with the initial condition $T_0 = 0$) are i.i.d. $\mathrm{Exp}(\lambda)$ random variables. Alternatively, if for all $0 \le a \le b$ we let

$$N(a,b) \triangleq \#\Big(\text{points } T_k \text{ in the interval } (a,b]\Big),$$

then $N(a,b) \sim \mathrm{Pois}(\lambda(b-a))$, and $N(a,b)$ is independent of $N(c,d)$ if $(a,b)\cap(c,d) = \varnothing$. Moreover, for any $r \in \mathbb{R}$ such that $a + r \ge 0$ and $b + r \ge 0$, $N(a+r, b+r)$ has the same distribution as $N(a,b)$. This is just another way of saying that the Poisson counting process

$$N_t = N(0,t) = \sum_{k=1}^{\infty} u(t - T_k), \tag{4.38}$$

where $u(\cdot)$ is the unit step function, has independent and stationary increments.

We can also associate each arrival with a unit impulse and form the stochastic signal

$$X_t = \sum_{i=1}^{\infty} \delta(t - T_i). \tag{4.39}$$

If $X = (X_t)_{t\ge 0}$ is an input to an LTI system with impulse response h, then the corresponding output is given by

$$Y_t = \sum_{i=1}^{\infty} h(t - T_i). \tag{4.40}$$

This type of stochastic signal is a model of *shot noise* in electronic and optical devices, which is a manifestation of the discrete nature of charge carriers or photons. In that context, h describes the

response of the device to individual arrivals. We will look at shot noise in more detail later; for now, we will state and prove a useful result, known as *Campbell's theorem*. The shot noise process Y has the mean

$$m_Y = \lambda \int_{-\infty}^{\infty} h(\tau) \mathrm{d}\tau = \lambda H(0) \tag{4.41}$$

and the variance

$$\mathrm{Var}[Y_t] = \lambda \int_{-\infty}^{\infty} h^2(\tau) \mathrm{d}\tau. \tag{4.42}$$

Let us first prove (4.41). Since

$$Y_t = \int_{-\infty}^{\infty} h(t-\tau) X_\tau \, \mathrm{d}\tau, \tag{4.43}$$

we can apply the formula (4.10):

$$\mathbf{E}[Y_t] = \mathbf{E}\left[\int_{-\infty}^{\infty} h(t-\tau) X_\tau \mathrm{d}\tau \right].$$

Interchanging the integral and the expectation, we can write

$$\mathbf{E}[Y_t] = \int_{-\infty}^{\infty} h(t-\tau) \mathbf{E}[X_\tau] \mathrm{d}\tau, \tag{4.44}$$

so now we need to compute the mean $m_X(\tau)$. Since $\frac{\mathrm{d}}{\mathrm{d}t} u(t) = \delta(t)$, we can view the stochastic signal X as the result of passing the Poisson counting process N through a differentiator:

$$X_t = \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{\infty} u(t - T_i) = \frac{\mathrm{d}}{\mathrm{d}t} N_t. \tag{4.45}$$

Since $\mathbf{E}[N_t] = \lambda t$, we have

$$\mathbf{E}[X_t] = \mathbf{E}\left[\frac{\mathrm{d}}{\mathrm{d}t} N_t \right] = \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{E}[N_t] = \lambda$$

(once again, we have been somewhat cavalier about interchanging the order of expectation and differentiation, but all of these steps can be made rigorous using tools from the theory of point processes). Substituting this into (4.44), we get (4.41). Now we prove (4.42). Using the definition of variance and (4.41), we have

$$\begin{aligned} \mathrm{Var}[Y_t] &= \mathbf{E}[Y_t^2] - (\mathbf{E}[Y_t])^2 \\ &= \mathbf{E}[Y_t^2] - \lambda^2 \left(\int_{-\infty}^{\infty} h(\tau) \mathrm{d}\tau \right)^2, \end{aligned} \tag{4.46}$$

so now we need to compute $\mathbf{E}[Y_t^2]$. Using (4.43), we can write

$$\begin{aligned}\mathbf{E}[Y_t^2] &= \mathbf{E}\left[\left(\int_{-\infty}^{\infty} h(t-\tau)X_\tau \mathrm{d}\tau\right)\left(\int_{-\infty}^{\infty} h(t-\tau')X_{\tau'}\mathrm{d}\tau'\right)\right] \\ &= \mathbf{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\tau)h(t-\tau')X_\tau X_{\tau'}\mathrm{d}\tau\mathrm{d}\tau'\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\tau)h(t-\tau')\mathbf{E}[X_\tau X_{\tau'}]\mathrm{d}\tau\mathrm{d}\tau'. \end{aligned} \tag{4.47}$$

Once again, we use (4.45):

$$\begin{aligned}\mathbf{E}[X_\tau X_{\tau'}] &= \mathbf{E}\left[\frac{\partial^2}{\partial\tau\partial\tau'}N_\tau N_{\tau'}\right] \\ &= \frac{\partial^2}{\partial\tau\partial\tau'}\mathbf{E}[N_\tau N_{\tau'}] \\ &= \frac{\partial^2}{\partial\tau\partial\tau'}R_N(\tau,\tau'). \end{aligned} \tag{4.48}$$

The autocorrelation function of the Poisson process with rate λ is given by

$$R_N(s,t) = \lambda^2 st + \lambda\min\{s,t\}$$

(exercise: verify this!), and therefore

$$\begin{aligned}\frac{\partial^2}{\partial s\partial t}R_N(s,t) &= \frac{\partial^2}{\partial s\partial t}\left(\lambda^2 st + \lambda\min\{s,t\}\right) \\ &= \frac{\partial}{\partial s}\left(\lambda^2 s + \lambda\frac{\partial}{\partial t}\min\{s,t\}\right).\end{aligned}$$

For a fixed value of s,

$$\min\{s,t\} = \begin{cases} s, & t > s \\ t, & t \le s \end{cases},$$

from which it follows that

$$\frac{\partial^2}{\partial s\partial t}\min\{s,t\} = \frac{\partial}{\partial s}u(s-t) = \delta(s-t).$$

Consequently,

$$\frac{\partial^2}{\partial s\partial t}R_N(s,t) = \lambda^2 + \lambda\delta(s-t).$$

Using this in (4.48) and (4.47), we have

$$\begin{aligned}\mathbf{E}[Y_t^2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\tau)h(t-\tau')\Big[\lambda^2+\lambda\delta(\tau-\tau')\Big]\mathrm{d}\tau\mathrm{d}\tau' \\ &= \lambda^2\left(\int_{-\infty}^{\infty} h(\tau)\mathrm{d}\tau\right)^2 + \lambda\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\tau)h(t-\tau')\delta(\tau-\tau')\mathrm{d}\tau\mathrm{d}\tau' \\ &= \lambda^2\left(\int_{-\infty}^{\infty} h(\tau)\mathrm{d}\tau\right)^2 + \lambda\int_{-\infty}^{\infty} h^2(\tau)\mathrm{d}\tau.\end{aligned}$$

Substituting this expression into (4.46), we obtain (4.42).

4.7 Markov chains and linear systems

Another probabilistic context in which linear systems arise is the evolution of the probability distribution of the state of a Markov chain. Specifically, given a Markov chain $X = (X_t)_{t\in\mathbb{Z}_+}$, if we denote by p_t the probability distribution of its state at time t, i.e., $p_t(x) = \mathbf{P}[X_t = x]$ for each x in the state space X, then

$$p_{t+1} = p_t M. \tag{4.49}$$

Here, $M = [M(x,y)]_{x,y\in\mathsf{X}}$ is the matrix of one-step transition probabilities of the chain.

As we already had seen through the example of the PageRank algorithm, a wide variety of algorithms in signal processing and machine learning can be viewed as iterated applications of the transformation (4.49). We will now see another example of such an algorithm, the so-called *average consensus* (or *linear agreement dynamics*). The idea is as follows. Suppose we have a finite number n of entities (e.g., robots or sensors in a sensor network). We will refer to these entities as *agents*. Label the agents by the elements the set $\mathsf{X} = \{1,\dots,n\}$. With each $x \in \mathsf{X}$, we associate a real number $f_0(x)$. Let us stack them into a *column vector* $f_0 = (f_0(1),\dots,f_0(n))^T$. We will think of $f_0(x)$ as the *initial observation* at x. Suppose that each agent can exchange real-valued messages with a subset of the agents. This defines a *communication network*. For example, Figure 2 shows a *ring network* with 5 agents. For each x, let $N(x)$ denote the set of *neighbors of* x in the network. We assume that the communication goes both ways, i.e., if $y \in N(x)$, then $x \in N(y)$. For example, in Figure 2, we have

$$N(1) = \{2,5\}, N(2) = \{1,3\}, N(3) = \{2,4\}, N(4) = \{3,5\}, N(5) = \{1,4\}.$$

The goal is for all the agents to learn the average value of everyone's observations,

$$a = \frac{1}{n}\sum_{x\in\mathsf{X}} f_0(x).$$

The average consensus algorithm accomplishes this as follows. We pick an $n\times n$ matrix $M = [M(x,y)]_{x,y\in\mathsf{X}}$ with the following properties:

1. Nonnegativity – $M(x,y) \geq 0$ for all x, y.

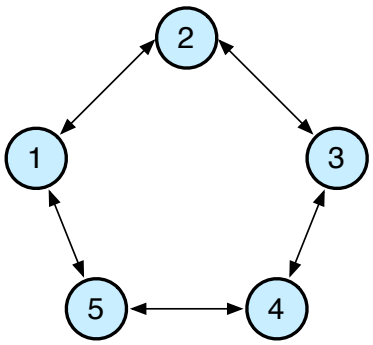

Figure 2: Ring network consisting of 5 agents.

2. Network conformity – $M(x,y) \neq 0$ only if $x = y$ or if $x \in N(y)$.

3. Double Markov property – $\sum_{y \in \mathsf{X}} M(x,y) = 1$ and $\sum_{x \in \mathsf{X}} M(x,y) = 1$ for all x, y.

The average consensus algorithm generates a sequence of column vectors $(f_t)_{t \in \mathbb{Z}_+}$ starting with the given initial condition via

$$f_{t+1} = M f_t. \tag{4.50}$$

We will see that, under certain conditions on M, each $f_t(x)$ will converge to a. Before getting to that, though, let us discuss some properties of the consensus algorithm.

First of all, note that, for each $x \in \mathsf{X}$,

$$\begin{aligned} f_{t+1}(x) &= \sum_{y \in \mathsf{X}} M(x,y) f_t(y) \\ &= M(x,x) f_t(x) + \sum_{y \in N(x)} M(x,y) f_t(y). \end{aligned}$$

Thus, $f_{t+1}(x)$ depends only on the coordinates of f_t held by agent x and by the neighobrs of agent x. Secondly, for each t define

$$a_t \triangleq \frac{1}{n} \sum_{x \in \mathsf{X}} f_t(x).$$

Then

$$a_t = a, \qquad \forall t = 0, 1, 2, \ldots. \tag{4.51}$$

Indeed, $a_0 = a$, and, using the double Markov property of M, we can write

$$\begin{aligned} a_{t+1} &= \frac{1}{n}\sum_{x\in\mathsf{X}} f_{t+1}(x) \\ &= \frac{1}{n}\sum_{x\in\mathsf{X}}\sum_{y\in\mathsf{X}} M(x,y) f_t(y) \\ &= \frac{1}{n}\sum_{y\in\mathsf{X}}\Big(\sum_{x\in\mathsf{X}} M(x,y)\Big) f_t(y) \\ &= \frac{1}{n}\sum_{y\in\mathsf{X}} f_t(y) \\ &= a_t. \end{aligned}$$

Thus, the consensus algorithm preserves averages.

With these two observations at hand, we will now show that, under certain conditions on M, the average consensus algorithm (4.50) converges. To quantify convergence, let us introduce the following function of an arbitrary column vector $f = (f(1),\dots,f(n))^T$:

$$V(f) \triangleq \frac{1}{n}\sum_{x\in\mathsf{X}} (f(x)-a)^2. \tag{4.52}$$

This function can be interpreted geometrically as follows: If we define the Euclidean norm of a vector f by

$$\|f\| = \sqrt{\sum_{x\in\mathsf{X}} f^2(x)}$$

then

$$V(f) = \frac{1}{n}\|f - ae\|^2,$$

where $e = (1,1,\dots,1)^T$ is the column vector of all ones. That is, $V(f)$ is the square of the distance between the points f and ae, normalized by the number of agents n. Clearly, $V(f) \geq 0$, and $V(f) = 0$ if and only if $f(x) = a$ for all x. Thus, if we can show that $V(f_t) \to 0$ as $t \to \infty$, then we will be able to conclude that each coordinate of f_t indeed converges to the average a. What is more, we will prove that, under certain conditions on M, $V(f_{t+1}) < V(f_t)$, unless $f_t = ae$, in which case $V(f_{t+1}) = V(f_t) = 0$.

To proceed, let us expand the difference $V(f_{t+1}) - V(f_t)$:

$$\begin{aligned} V(f_{t+1}) - V(f_t) &= \frac{1}{n}\Big(\|f_{t+1} - ae\|^2 - \|f_t - ae\|^2\Big) \\ &= \frac{1}{n}\sum_{x\in\mathsf{X}}\Big(f_{t+1}^2(x) - 2a f_{t+1}(x) + a^2 - f_t^2(x) + 2a f_t(x) - a^2\Big) \\ &= \frac{1}{n}\sum_{x\in\mathsf{X}}\Big(f_{t+1}^2(x) - f_t^2(x)\Big) + \frac{2a}{n}\sum_{x\in\mathsf{X}}(f_t(x) - f_{t+1}(x)). \end{aligned} \tag{4.53}$$

Now, the second term in (4.53) is equal to $a(a_{t+1} - a_t)$, which is equal to zero by (4.51). The first term, on the other hand, is equal to

$$\frac{1}{n}\left(\|f_{t+1}\|^2 - \|f_t\|^2\right) = \frac{1}{n}\left(\|Mf_t\|^2 - \|f_t\|^2\right).$$

Thus, we obtain the formula

$$V(f_{t+1}) - V(f_t) = \frac{1}{n}\|Mf_t\|^2 - \frac{1}{n}\|f_t\|^2, \tag{4.54}$$

and therefore we are led to study the relationship between the norms $\|Mf\|$ and $\|f\|$ for an arbitrary column vector f. First, we note that $Me = e$. Indeed,

$$(Me)(x) = \sum_{y \in \mathsf{X}} M(x,y) = 1.$$

Thus, if $f = ce$ for any $c \in \mathbb{R}$, then $Mf = f$, and so $\|Mf\| = \|f\|$. Secondly, a simple calculation shows that

$$\|Mf\|^2 - \|f\|^2 = \left\|M\left(f - \frac{f^T e}{n} e\right)\right\|^2 - \left\|f - \frac{f^T e}{n} e\right\|^2. \tag{4.55}$$

Thus, we can assume, without loss of generality, that $f^T e = 0$. Define the following quantity:

$$\gamma(M) \triangleq \min \frac{\|f\|^2 - \|Mf\|^2}{\|f\|^2}, \tag{4.56}$$

where the minimum is over all nonzero column vectors f that satisfy $f^T e = 0$. This quantity is referred to as the *spectral gap* of M, for reasons that will be explained later on. Now let us get back to the consensus algorithm. Suppose that $f_t \neq ae$. Then the vector $g_t = f_t - ae$ is nonzero and satisfies

$$\begin{aligned} g_t^T e &= \sum_{x \in \mathsf{X}} g_t(x) \\ &= \sum_{x \in \mathsf{X}} f_t(x) - na \\ &= n(a_t - a) \\ &= 0. \end{aligned}$$

Therefore, using the definition of $\gamma(M)$ in (4.56) we can write

$$\begin{aligned} V(f_{t+1}) - V(f_t) &= \frac{1}{n}\left(\|Mf_t\|^2 - \|f_t\|^2\right) \\ &= \frac{1}{n}\left(\|M(f_t - ae)\|^2 - \|f_t - ae\|^2\right) \\ &\leq -\gamma(M)\frac{1}{n}\|f_t - ae\|^2 \\ &= -\gamma(M)V(f_t). \end{aligned}$$

Rearranging, we obtain

$$V(f_{t+1}) \le (1-\gamma(M))V(f_t). \tag{4.57}$$

We will now use this inequality to analyze the convergence of average consensus.

To start with, we claim that $\gamma(M) \in [0,1]$, which implies that $V(f_{t+1}) \le V(f_t)$ for each t. To see this, let f be an arbitrary column vector. Then

$$\|Mf\|^2 = \sum_{x\in\mathsf{X}} \Big| \sum_{y\in\mathsf{X}} M(x,y)f(y) \Big|^2. \tag{4.58}$$

Recall the *Cauchy–Schwarz inequality*: for any k pairs of real numbers $(a_1,b_1),\ldots,(a_k,b_k)$,

$$\Big| \sum_{i=1}^k a_i b_i \Big| \le \sqrt{\sum_{i=1}^k a_i^2 \sum_{j=1}^k b_j^2}. \tag{4.59}$$

Then, for any $x \in \mathsf{X}$,

$$\begin{aligned} \Big| \sum_{y\in\mathsf{X}} M(x,y)f(y) \Big|^2 &= \Big| \sum_{y\in\mathsf{X}} \sqrt{M(x,y)} \cdot \sqrt{M(x,y)} f(y) \Big|^2 \\ &\le \sum_{y\in\mathsf{X}} M(x,y) \cdot \sum_{y\in\mathsf{X}} M(x,y) f^2(y) \\ &= \sum_{y\in\mathsf{X}} M(x,y) f^2(y), \end{aligned}$$

where the second step is by (4.59), while the third line uses the Markov property of M. Substituting this into (4.58), we obtain

$$\begin{aligned} \|Mf\|^2 &\le \sum_{x\in\mathsf{X}} \sum_{y\in\mathsf{X}} M(x,y) f^2(y) \\ &= \sum_{y\in\mathsf{X}} \Big(\sum_{x\in\mathsf{X}} M(x,y) \Big) f^2(y) \\ &= \sum_{y\in\mathsf{X}} f^2(y) \\ &= \|f\|^2, \end{aligned}$$

where we have used the symmetry and the Markov property of M. This shows that $\|Mf\|^2 \le \|f\|^2$ for any f. Therefore, $0 \le \|f\|^2 - \|Mf\|^2 \le \|f\|^2$, and so $\gamma(M) \in [0,1]$, as claimed. This shows that each iteration of the average consensus algorithm brings the coordinates of f_t closer to the average a, but it does not imply that f_t converges to ae as $t \to \infty$. To show that, we need stronger conditions that have to do wtih the *eigenvalues* of M^2 (the square of M). Recall that the eigenvalues of a square matrix A are the roots of the characteristic polynomial

$$\det(A - \lambda I) = 0,$$

where I is the identity matrix of the same shape as A, and that any vector f that satisfies $Af = \lambda f$ is an eigenvector of A with eigenvalue λ. The set of all eigenvalues of A is called the *spectrum* of A. Since any degree-n polynomial with real coefficients has n complex roots (possibly repeated), the spectrum of A is a finite subset of $\mathbb{C}$. However, if A is symmetric, i.e., $A = A^T$, then it has only real eigenvalues. Moreover, if A is positive, i.e., if $f^T Af \geq 0$ for all vectors f, then all eigenvalues of A are nonnegative.

Let us now apply these considerations to the matrix M^2. This matrix is obviously symmetric (because M is), and it is also positive. Indeed, for any column vector f we have

$$f^T M^2 f = (Mf)^T Mf = \|Mf\|^2 \geq 0.$$

Hence, all eigenvalues of M^2 are real and nonnegative. In fact, the spectrum of M^2 is contained in the unit interval $[0,1]$. To see this, suppose that λ is a nonzero eigenvalue of M^2, with eigenvector $f \neq 0$. Then

$$\|f\|^2 \geq \|Mf\|^2 = \lambda^2 \|f\|^2,$$

which gives $\lambda^2 \leq 1$. Since $\lambda \geq 0$, it must be the case that $\lambda \in [0,1]$. Moreover, $\lambda = 1$ is an eigenvalue: since $Me = e$, we have $M^2 e = M(Me) = Me = e$. Now we are finally ready to state the following key result:

> Suppose that M is such that any eigenvector of M with eigenvalue 1 is proportional to e. Then $1 - \gamma(M)$ is equal to the second largest eigenvalue of M^2, and consequently $1 - \gamma(M) < 1$.

To see why $\gamma(M)$ is related to the eigenvalues of M^2 in this way, let f be any of its eigenvectors with eigenvalue $\lambda < 1$. We know that f is not proportional to e, by assumption. What is more, we claim that $c = f^T e = 0$. Suppose not. Note that we can write $f = \frac{c}{n} e + g$, where $g = f - \frac{c}{n} e$. Then $g^T e = f^T e - \frac{c}{n} e^T e = c - c = 0$, i.e., g is orthogonal to any nonzero vector proportional to e. By linearity and the fact that $M^2 e = e$, we have

$$\begin{aligned} M^2 f &= M^2 \left(\frac{c}{n} e + g \right) \\ &= \frac{c}{n} M^2 e + M^2 g \\ &= \frac{c}{n} e + M^2 g. \end{aligned}$$

On the other hand, we have

$$M^2 f = \lambda f = \frac{\lambda c}{n} e + \lambda g$$

This implies that $\frac{c}{n} e + M^2 g = \frac{\lambda c}{n} e + \lambda g$. Since $g^T e = 0$ and $(M^2 g)^T e = g^T M^2 e = g^T e = 0$, it follows that $\lambda = 1$, which is a contradiction unless $c = 0$. Hence, if $M^2 f = \lambda f$ with $\lambda < 1$, then $f^T e = 0$. Therefore,

$$\gamma(M) \leq \frac{\|f\|^2 - \|Mf\|^2}{\|f\|^2} = 1 - \lambda,$$

or, equivalently, $\lambda \le 1 - \gamma(M)$. On the other hand, let λ^* be the second largest eigenvalue of M^2. Then any nonzero vector f with $f^T e = 0$ must satisfy the inequality $\|Mf\|^2 \le \lambda^* \|f\|^2$ (why?), which would imply that

$$\gamma(M) = \min\left\{ \frac{\|f\|^2 - \|Mf\|^2}{\|f\|^2} : f \neq 0, f^T e \right\} \ge 1 - \lambda^*.$$

Since we also have $1 - \lambda^* \le \gamma(M)$, we conclude that $1 - \gamma(M) = \lambda^* < 1$. Now we see why we have called $\gamma(M)$ the spectral gap: the spectrum (i.e., the set of all eigenvalues) of M^2 is contained in the set $[0, 1 - \gamma(M)] \cup \{1\}$, and therefore $\gamma(M)$ is the gap between the largest and the second largest eigenvalues of M^2. Thus, if M has the spectral gap property,

$$V(f_t) \le (1 - \gamma(M)) V(f_{t-1}) \le \ldots \le (1 - \gamma(M))^t V(f_0),$$

where $1 - \gamma(M) \in [0, 1)$. Thus, as $t \to \infty$,

$$V(f_t) = \frac{1}{n} \|f_t - ae\|^2 \to 0,$$

and the convergence is exponentially fast. This type of analysis occurs quite often in the context of iterative algorithms, and can be summarized as follows:

- We have a function $V(f)$ that assigns a nonnegative number to each column vector f, and is equal to zero if and only if $f = ae$.
- $V(f_t) = \frac{1}{n}\|f_t - ae\|^2$ is a measure of *progress* of the algorithm towards the goal of each agent having an accurate estimate of the average $a = \frac{1}{n} f_0^T e$.
- Each iteration of the algorithm strictly decreases V: $V(f_{t+1}) \le (1 - \gamma) V(f_t)$, where $1 - \gamma$ is a nonnegative constant strictly smaller than 1.

Then it follows that, as we run more and more iterations, the vectors f_t converge to ae, which corresponds to the ideal situation when each agent knows the average value a exactly. In general, if f_0 is not proportional to e, then $f_t \neq ae$ for any t (unless $\gamma(M) = 1$), but, as $t \to \infty$, $|f_t(x) - a| \to 0$ exponentially fast for each $x \in \mathsf{X}$. This method is referred to as the *method of Lyapunov* (or *potential*) *functions* — in many cases where we wish to prove that the state f_t of some iterative algorithm converges to some desired steady state f_*, we can find a function $V(f)$ that has the same three properties as above, with f_* in place of ae. Then we say that V is the Lyapunov function for the algorithm under consideration, and then it can be proved that f_t indeed converges to f_*.

Of course, verifying the spectral gap property is not straightforward. Here is one useful criterion, called the *Perron–Frobenius theorem*:

> Let A be an $n \times n$ matrix with the following properties:
>
> - the entries of A are nonnegative;
> - for any nonzero vector $f = (f(1), \ldots, f(n))^T$ with all $f(x) \ge 0$, the vector Af is positive, i.e., $Af(x) > 0$ for all x;

- $\lambda = 1$ is an eigenvalue of A, with a positive eigenvector f_*.

Then any other eigenvector of A with eigenvalue 1 is a multiple of f_*.

For example, if all entries of A are positive, then the second property will hold. Also, it is not hard to see that, even if A does not satisfy these conditions but some power A^k does, then the conclusion will still hold (why?).

Let us get back to our setting of communication between agents in a network. Suppose first that each agent has the same number of neighbors, i.e., $|N(x)| = |N(y)|$ for all x, y (in this case, we say that the network graph is regular). Then we can take the matrix

$$M(x,y) = \begin{cases} \frac{1}{d+1}, & \text{if } x = y \text{ or } y \in N(x) \\ 0, & \text{otherwise} \end{cases},$$

where d is the number of neighbors of any agent. Let us verify that this matrix has the required properties. First of all, each $M(x,y)$ is obviously nonnegative. Next, we see that $M(x,y)$ is zero only if $x \neq y$ and $y \notin N(x)$. Finally, we want to establish the double Markov property. First, for any $x \in \mathsf{X}$ we have

$$\begin{aligned} \sum_{y \in \mathsf{X}} M(x,y) &= M(x,x) + \sum_{y \in N(x)} M(x,y) \\ &= \frac{1}{d+1} + \sum_{y \in N(x)} \frac{1}{d+1} \\ &= \frac{1}{d+1} + \frac{d}{d+1} \\ &= 1. \end{aligned}$$

Likewise, for any $y \in \mathsf{X}$,

$$\begin{aligned} \sum_{x \in \mathsf{X}} M(x,y) &= M(y,y) + \sum_{x:\, y \in N(x)} M(x,y) \\ &= \frac{1}{d+1} + \sum_{x:\, y \in N(x)} \frac{1}{d+1}. \end{aligned}$$

So, we have a summation over all x that have a given y as their neighbor. Since x is a neighbor of y if and only if y is a neighbor of x, we have

$$\sum_{x:\, y \in N(x)} \frac{1}{d+1} = \sum_{x:\, x \in N(y)} \frac{1}{d+1} = \frac{d}{d+1},$$

and therefore $\sum_{x \in \mathsf{X}} M(x,y) = 1$. For example, for the network shown in Fig. 2 we have

$$M = \begin{pmatrix} 1/3 & 1/3 & 0 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 0 & 1/3 & 1/3 \end{pmatrix}.$$

The matrix M^2 has eigenvalues 1, $\frac{3+\sqrt{5}}{18}$ (with multiplicity 2), and $\frac{3-\sqrt{5}}{18}$ (with multiplicity 2), and therefore

$$\gamma(M) = 1 - \frac{3+\sqrt{5}}{18} = \frac{15+\sqrt{5}}{18}.$$

If the network graph is not regular, then the matrix

$$M = \begin{cases} 1 - \frac{|N(x)|}{d_{\max}+1}, & \text{if } x = y \\ \frac{1}{d_{\max}+1}. & \text{if } x \in N(y)\,, \\ 0, & \text{otherwise} \end{cases}$$

where $d_{\max} \triangleq \max_{x\in\mathsf{X}} |N(x)|$, has the required properties (prove it).

Just like the PageRank algorithm, the average consensus algorithm is deterministic. However, because the matrix M can be seen as a transition matrix of a Markov chain with state space $\mathsf{X} = \{1,\dots,n\}$, we can give an appealing stochastic interpretation of our analysis. Let $\pi = (1/n, 1/n, \dots, 1/n)$ denote the uniform probability distribution on X. Then $\pi = \pi M$, i.e., π is the invariant distribution of M. Indeed, using the double Markov property of M, we can write

$$\pi M(x) = \sum_{y\in\mathsf{X}} \pi(y) M(y,x) = \frac{1}{n} \sum_{y\in\mathsf{X}} M(y,x) = \frac{1}{n}.$$

Now, suppose that we have a Markov chain $X = (X_t)_{t\in\mathbb{Z}_+}$ with this transition matrix, and we have some real-valued function f_0 on the state space X. For a given initial state $X_0 = x_0$, let us calculate the expected value of $f_0(X_t)$. It is given by

$$\sum_{x\in\mathsf{X}} M^t(x_0,x) f_0(x) = f_t(x_0).$$

This is a function of x_0. If $X_0 \sim \pi$, then this becomes a random variable whose mean is given by

$$\mathbf{E}[f_t(X_0)] = \sum_{x\in\mathsf{X}} \pi(x) f_0(x) = \frac{1}{n} \sum_{x\in\mathsf{X}} f_0(x) = a,$$

and whose variance is

$$\mathrm{Var}[f_t(X_0)] = \sum_{x\in\mathsf{X}} \pi(x) \left(f_t(x) - a\right)^2 = \frac{1}{n} \sum_{x\in\mathsf{X}} \left(f_t(x) - a\right)^2 = \frac{1}{n} \|f_t - ae\|^2 = V(f_t).$$

Thus, our analysis of the average consensus algorithm can be viewed in terms of Markov chains: Suppose that we have a Markov chain on X with transition matrix M which is double Markov, and such that π is an invariant distribution. Suppose also that the column vector $e = (1,1,\dots,1)^T$ is the only eigenvector of M^2 with eigenvalue 1. Then, if we start the chain in the invariant distribution, i.e., $X_0 \sim \pi$, then, for any function $f_0 : \mathsf{X} \to \mathbb{R}$,

$$\mathrm{Var}[f_t(X_0)] \le (1 - \gamma(M))^t \mathrm{Var}[f_0(X_0)],$$

where $\gamma(M) \in [0,1)$ is the spectral gap of M.