Universidad de la República
Centro de Matemática

Limit theorems for continuous time Markov chains and applications to large scale queueing systems

Tesis de Maestría en Matemática

Diego Goldsztajn[1]
diegogolds@gmail.com

Orientadores:
Fernando Paganini - paganini@ort.edu.uy
Andrés Ferragut - ferragut@ort.edu.uy

Octubre 2018

[1]Becario del Sistema Nacional de Becas de la ANII, código POS_NAC_2016_1_130333.

A mis padres, mi hermana y mi novia por apoyarme, contenerme y motivarme en todo momento.

A Fernando y Andrés por todo lo que me han enseñado y porque es un placer trabajar con ellos.

Abstract

This thesis discusses limit theorems for density dependent families of continuous time Markov chains and their application to the stochastic analysis of large scale cloud computing environments and data centers. On the purely theoretical side, we review the classic functional strong law of large numbers and central limit theorem due to Kurtz, which characterize the asymptotic behavior of density dependent families in terms of their drift. In the case of the central limit theorem we provide extensions in two directions: to consider small order perturbations in the transition rates of the family and non-differentiable drifts. The classic theorems and the latter extensions are used to study the dynamic right sizing of capacity in large scale cloud environments and data centers, aimed at the adjustment of this capacity to an uncertain workload. Under a central queue scheme and Markovian assumptions, we design a policy that eliminates queueing almost completely, at the expense of a slight over-provisioning; if ρ the traffic intensity, then the over-provisioning scales as $O(\sqrt{\rho})$ when $\rho \to \infty$. In this sense our policy automatically adjusts the system's capacity according to the well-known square root staffing rule.

Key words: Markov chain, strong law of large numbers, fluid limit, central limit theorem, diffusion approximation, queueing theory, heavy traffic, feedback control, cloud computing, data center, auto-scaling.

Resumen

En esta tesis se estudian teoremas límite para familias de cadenas de Markov de tiempo continuo, así como su aplicación al análisis estocástico de ambientes tipo cloud y data centers. En un comienzo se presentan resultados clásicos debidos a Kurtz, que caracterizan el comportamiento asintótico de estas familias a partir de su drift; a saber, una ley fuerte de grandes números y un teorema central del límite, ambos funcionales. En el último caso obtenemos extensiones en dos direcciones: considerando perturbaciones de pequeño orden en las tasas de transición de la familia y drifts no diferenciables. Los teoremas clásicos y las extensiones anteriores se emplean para estudiar el ajuste dinámico de la capacidad de cómputo de ambientes tipo cloud y data centers de gran escala, orientado a ajustar la capacidad de cómputo a una demanda incierta. Utilizando un esquema de cola centralizada y bajo hipótesis Markovianas, diseñamos una política que evita el encolado de tareas a expensas de un pequeño sobre dimensionamiento de la capacidad de cómputo; si ρ is la intensidad de tráfico, entonces la capacidad ociosa escala como $O(\sqrt{\rho})$ cuando $\rho \to \infty$. En este sentido nuestra política ajusta automáticamente la capacidad de cómputo del sistema según el conocido criterio de la raíz cuadrada.

Palabras clave: cadena de Markov, ley furte de los grandes números, límite fluido, teorema central del límite, difusión, teoría de colas, control automático, heavy traffic, computación en la nube, data center, auto-scaling.

Contents

Introduction

In spite of their simple structure, Markov chains can be used to describe the behavior of a wide variety of random phenomena evolving in time, and because of this they have been extensively used in applied probability along the years. For instance, in the study of epidemics, a vector valued Markov chain can be used to describe how the number of infected and immune people evolve over time. Another example, from the field of chemistry, is the study of chemical reactions, where the state space of the Markov chain has vectorial nature as well, and the coordinates represent the amount of reactants and products at a given time.

An application that is more relevant to the scope of this thesis is the study of computing systems, which belong to the much broader class of queueing systems; the reader may find a classical study of the latter objects in [18, 19]. Within the framework of queueing theory, the simplest model of a computing system is a first-come-first-served queue with a single server. In this model, requests requiring to perform a certain task or job arrive sequentially to the system and are stored in the queue, where they wait to be processed at the server, in order of arrival; some relevant parameters are the arrival rate of job requests, the service rate of jobs and the quotient of these two, the traffic intensity or workload that the system faces. Considering a number of servers that is greater than one, we may model the behavior of a larger class of computing systems, and if we moreover let the number of servers change over time, then we may study the behavior of modern data centers and cloud computing environments.

Under suitable hypothesis, the behavior of computing systems may be described using continuous time Markov chains. Usually, the relevant questions about the performance of computing systems concern their typical or stationary behavior, and in the Markovian framework these questions can be formulated in terms of the invariant distribution of the chain. Unfortunately, an explicit computation of this distribution is usually not possible, especially when the dimension of the state space is higher than one; solving the balance equations of the Markov chain is in general prohibitively involved. This is the situation when the number of servers changes over time, here the state of the chain must store information about both the number of tasks and servers in the system and thus we have a bidimensional state-space.

In order to overcome this hurdle, a standard methodology is to let the arrival rate of jobs approach infinity, after an adequate normalization this results in a sequence of Markov chains that converges to a process which is sometimes easier to analyze; some

of the first works to embrace this approach are [5, 15]. This procedure is especially justified in the large scale context of modern data centers: with facilities that may reach the 60 Hectare of size, the equivalent of 110 football pitches. Depending on the type of normalization that we adopt, the result is a law of large numbers or a central limit theorem. In the first case, the limit process is deterministic and solves an ordinary differential equation (ODE), which arises naturally from the transition rates of the Markovian model. This sheds light on the macroscopic behavior of the computing system, nevertheless it removes all stochasticity, which warrants taking a closer look. To achieve this, we adopt a different normalization and in this case the process that we see after taking the limit is a diffusion, that solves a stochastic differential equation (SDE). The stationary distribution of this process may be used to estimate the typical behavior of the system that we are modeling. The quality of the approximations that derive from these limit procedures may be judge by numerical comparisons, but in any case this methodology provides a valuable insight on the asymptotic behavior of the metrics that characterize the system, and their relative orders of magnitude.

The technical name for the sequences of continuous time Markov chains that arise when we consider increasing arrival rates approaching infinity is density dependent families, and the classical limit theorems in this setting are due to Kurtz; the reader may find them in [8, 20–22]. In this work we review these theorems, providing detailed proofs, with the intention of using them in the analysis of computing systems where the number of servers is being dynamically right sized for an improved performance. Nevertheless, the fact that some of these systems do not fit the hypothesis of the latter theorems motivates us to develop extensions, particularly in the case of the central limit theorem. Afterwards, we propose feedback control rules to right size the capacity of computing systems, and use these extensions to assess their performance, elucidating the minimum over-provisioning, in terms of idle servers, that ensures virtually none queueing delay to customers.

Contributions of this work

In the classical theorems due to Kurtz, the limit behavior of density dependent families is characterized by their drift: a vector field that is constructed using the intensities of the chains in the family. Indeed, the limit process in the law of large numbers, called fluid limit, is deterministic and solves an ODE whose field is the drift. Moreover, in the central limit theorem the limit is a diffusion, which solves a SDE that may be written in terms of the drift's Jacobian matrix.

Naturally, the hypothesis of the last theorem require the drift of the family to be smooth. One of the contributions of this work is a central limit theorem, around an equilibrium point of the fluid dynamics, for density dependent families whose drift is not differentiable at the latter point. This is an important result, because central limit theorems around globally asymptotically stable fluid equilibriums are especially useful for estimating the steady-state behavior of computing systems.

Another contribution of this work is an extension of the classical theorems to families whose elements may display small order perturbations in their transition rates. Kurtz had already considered perturbed intensities in [22], but in his work the perturbations disappear both in the fluid and diffusion scale. The kind of perturbations that we consider in this work disappear in the fluid scale as well, however they give rise to a new term in the SDE that appears in the diffusion scale.

Finally, this work proposes a feedback control rule designed to right size the capacity of computing systems with a centralized queue. Denoting by ρ the traffic intensity that the system faces, this rule achieves virtually zero queueing delay at the expense of an average over-provisioning of $O(\sqrt{\rho})$ idle servers. Thus, our rule automatically tracks the Halfin-Whitt regime, which was originally described in [14].

Organization of the thesis

We begin Chapter 1 specifying the mathematical model of a computing system that we will use throughout the thesis. Afterwards, we use a traditional example, the infinite-server queue, to illustrate the model and also some of the ideas that give birth to the theorems of the following chapter.

In Chapter 2 we define density dependent families of continuous time Markov chains and we review the classical limit theorems due to Kurtz. Here we provide a detailed proof of the law of large numbers, but we do not give a full proof of the central limit theorem to avoid repeating some of the arguments that will appear in the next chapter, where we provide extensions to this theorem. However, we outline the proof of the central limit theorem due to Kurtz, and we point out some differences with respect to the proofs that will appear in Chapter 3.

In Chapter 3 we extend the central limit theorem of Chapter 2 in two different directions, as it was described in the contributions section of this introduction. We provide full proofs of these theorems, and afterwards we discuss the problem of finding the stationary distribution of some switched diffusions, which arise when we consider the limit of a density dependent family whose drift is not differentiable.

The application of the latter theorems is illustrated in Chapter 4. Here we consider the problem of right sizing the capacity of computing systems and we provide several control rules with this objective in mind. The latter policies ultimately derive in the rule that was announced in the contributions section, which automatically tracks the Halfin-Whitt regime.

Finally, conclusions appear in Chapter 5, and additional material for the reader is provided in the appendices. Namely, Appendix A is concerned with the topology of Skorohod spaces and the weak convergence of processes that take values there. Appendix B contains limit theorems for the Poisson process. Appendix C refers to Markov processes and their characterization by means of semigroups and their infinitesimal generators. Appendix D concerns Itô calculus and stochastic differential equations. Finally, Appendix E contains the proofs of some useful propositions.

Comments on notation

In general we will consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we will use the term almost sure instead of the measure theoretic terminology $\mathbb{P}$-almost everywhere. Whenever we say that some property holds almost everywhere, this will mean that the property holds outside of a subset of $\mathbb{R}^d$ that is null with respect to the Lebesgue measure, in the sense that it is contained in a set of measure zero. In addition, we will adopt the following notation.

$\xrightarrow{\text{a.s.}}$	almost sure convergence
$\xrightarrow{\mathbb{P}}$	convergence in probability
$\xrightarrow{L^p}$	convergence in $L^p(\Omega)$
$\Rightarrow$	convergence in distribution
$\mathbb{E}$	expectation
$\mathbb{V}$	variance
$\sigma(\mathcal{A})$	σ-algebra generated by $\mathcal{A}$

We will usually consider stochastic processes on $\mathbb{R}^d$. The symbol $||\cdot||$ will denote any norm in $\mathbb{R}^d$, the choice of the norm will not matter in general; in the few cases where it matters we will specify the norm. We will further use the next notation.

$B(\mathbb{R}^d)$	measurable and bounded real functions
$C(\mathbb{R}^d)$	continuous real functions
$C_b(\mathbb{R}^d)$	continuous and bounded real functions
$C_0(\mathbb{R}^d)$	continuous real functions that vanish at infinity
$C_c^k(\mathbb{R}^d)$	k times continuously differentiable real functions with compact support
$C_c^\infty(\mathbb{R}^d)$	infinitely differentiable real functions with compact support
x^+	$\max(x, 0)$

Notation

$q^k_{x,y} = k\beta^k_{k(y-x)}(x)$.

$\beta^k_l(x) = \gamma_l(x) + \delta^k_l(x)$.

$F(x) = \sum_{l\in D} l\gamma_l(x)$ (drift).

$G_k(x) = \sum_{l\in D} l\delta^k_l(x)$ (perturbing drift).

$\Sigma_k(t) = \sum_{l\in D} \frac{l}{k} Y_l \left(\int_0^t k\beta^k_l(X_k(\tau))d\tau \right).$

$X_k(t) = X_k(0) + \Sigma_k(t) + \int_0^t F(X_k(\tau))d\tau + \int_0^t G_k(X_k(\tau))d\tau.$

$x(t) = x(0) + \int_0^t F(x(\tau))d\tau$ (fluid limit).

$Z_k(t) = \sqrt{k}[X_k(t) - x(t)]$.

$U_k(t) = \sqrt{k}\Sigma_k(t)$.

$\delta_k(t) = \int_0^t \sqrt{k}\left[G_k(X_k(\tau)) + R_\tau(X_k(\tau))\right] d\tau.$

$Z_k(t) = Z_k(0) + U_k(t) + \delta_k(t) + \int_0^t \partial F(Z_k(\tau))d\tau.$

$U(t) = \sum_{l\in D} lW_l \left(\int_0^t \gamma_l(x(\tau))d\tau \right).$

$Z(t) = Z(0) + U(t) + \int_0^t \partial F(Z(\tau)) + G(x(\tau))d\tau.$

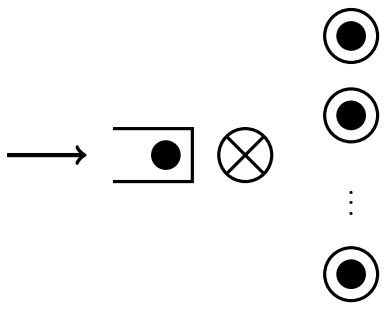
$dZ_t = [\partial F(Z_t) + G(x(t))]dt + B_t dW_t$ (diffusion approximation).

Chapter 1

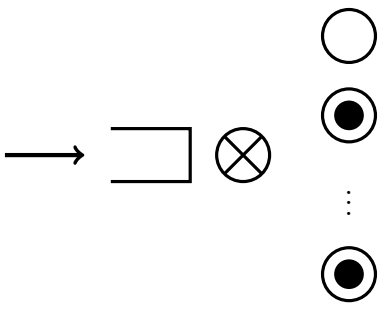
Modeling computing systems

1.1 Queueing system model

The model of a computing system that we will adopt falls into the framework of queueing theory, it comprises a single dispatcher, with a centralized queue, and a pool of server instances; this is illustrated in Figure 1.1. In this model jobs arrive to the dispatcher sequentially, where they are sent to an idle server unless all servers are busy; in the latter case jobs are queued in order of arrival and wait until some server becomes available. After receiving service, jobs leave the system.



(a) Since all servers are busy, jobs must be queued at the dispatcher.



(b) Arriving jobs are immediately dispatched to one of the idle servers.

Figure 1.1: Model of a computing system consisting of a single dispatcher with a centralized queue and a pool of servers. The queue is represented as a rectangular shape, the dispatcher is the crossed circle, servers are the white circles and jobs are depicted as black circles.

In this work job arrivals will be triggered by a Poisson process of intensity λ jobs per second, equivalently inter-arrival times will be independent and exponentially distributed with mean $1/\lambda$ seconds. The service time of jobs is the processing time that they require from servers, these times will be assumed to be independent and exponentially distributed as well, with mean $1/\mu$ seconds. A relevant parameter is the traffic intensity or workload that the system faces, which is defined as the ratio $\rho = \lambda/\mu$ between the mean service time and the mean inter-arrival time.

1.2 The infinite-server queue

In order to illustrate the methodology that we will use, we begin with a simple example: the infinite-server queue; in this queuing system a dedicated server is summoned upon the arrival of a new query, and this server leaves the system after processing the new job. Hence, the number of servers always matches the number of requests, and in particular there is no need to maintain a queue. Since the number of servers and jobs is the same, the behavior of this system can be characterized using the birth-death process depicted in Figure 1.2, which describes the evolution of the number of jobs in the system, or equivalently the number of servers. Note that the service time of a single job is exponential of parameter μ, thus when there are n tasks in the system, the time until one of them is finished is the minimum of n exponentials of parameter μ, which is exponential of parameter $n\mu$.

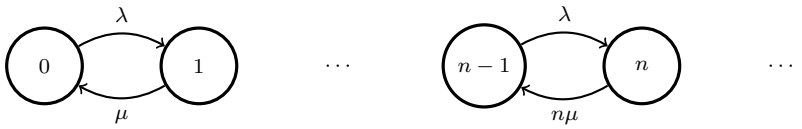


Figure 1.2: Birth-death process describing the number of jobs in an infinite-server queue.

Standard computations show that this birth-death process is ergodic. Furthermore, after writing the balance equations of the chain, it is easy to check that the stationary distribution π is Poisson of parameter ρ. In other words, the steady-state probability that there are exactly n jobs in the system is

$$\pi(n) = \frac{\rho^n e^{-\rho}}{n!} \quad \forall\ n \geq 0.$$

In particular, the mean and variance of the number of jobs both are equal to ρ in the steady-state.

The stationary distribution of the infinite-server queue can be easily computed from the chain's balance equations. However, in many other cases this is not possible and we must resort to an asymptotic analysis by means of limit theorems. Moreover, even in the present example, the limit theorems that we will see may help us characterize the transient behavior of the system.

1.2.1 Strong law of large numbers

In order to derive these theorems, in the context of the infinite-server queue, consider two independent Poisson processes with unitary intensity $\mathcal{N}_a$ and $\mathcal{N}_d$, defined over the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Also, consider a deterministic initial condition $X(0) \geq 0$, and let X be a process such that

$$X(t) = X(0) + \mathcal{N}_a(\lambda t) - \mathcal{N}_d\left(\int_0^t \mu X(\tau) d\tau\right) \quad \forall\ t \geq 0. \tag{1.1}$$

Such a process exists, is Markov and also unique, as it is shown in [8, Chapter 6.4]. Furthermore, when the system's initial occupation is $X(0)$ jobs, this process has the same infinitesimal generator as the chain represented in Figure 1.2; this fact is also proved in [8, Chapter 6.4] and is in line with one of the three characterizations of continuous time Markov chains that are given in [29, Chapter 2.6].

An interpretation of equation (1.1) is that the process $\mathcal{N}_a$ triggers job arrivals, while the process $\mathcal{N}_d$ represents departures from the system; note that the latter process can only increase, to indicate that a departure occurs, when the number of pending requests $X(t)$ is greater than zero.

It is convenient to consider the centered processes $Y_i(t) = \mathcal{N}_i(t) - t$, rather than the processes $\mathcal{N}_i$ themselves. We will also consider the field $F : \mathbb{R} \longrightarrow \mathbb{R}$ such that

$$F(x) = \lambda - \mu x, \tag{1.2}$$

which represents the mean drift away from the state x in the chain of Figure 1.2; not in vain this map is referred to as the drift of the chain. Using these new definitions, we may rewrite equation (1.1) as follows.

$$X(t) = X(0) + Y_a(\lambda t) - Y_d\left(\int_0^t \mu X(\tau) d\tau\right) + \int_0^t F\left(X(\tau)\right) d\tau \quad \forall\ t \geq 0. \tag{1.3}$$

Hence, X may be regarded as the solution to a stochastically perturbed version of the initial value problem $\dot{x} = F(x)$. In the large scale, when the traffic intensity ρ approaches infinity, and under an adequate normalization, we will see that the stochastic perturbations vanish.

To this purpose, consider a scale parameter $k \geq 1$ and a sequence of infinite-server queues $\hat{X}_k$, each with job arrival rate $k\lambda$; we are keeping the service rate μ fixed, and thus the workload $k\rho$ approaches infinity. All these processes may be constructed over $(\Omega, \mathcal{F}, \mathbb{P})$ in such a way that they satisfy the equations

$$\begin{aligned}\hat{X}_k(t) &= \hat{X}_k(0) + Y_a(k\lambda t) - Y_d\left(\int_0^t \mu \hat{X}_k(\tau) d\tau\right) + \int_0^t k\lambda - \mu \hat{X}_k(\tau) d\tau \\ &= \hat{X}_k(0) + Y_a(k\lambda t) - Y_d\left(\int_0^t \mu \hat{X}_k(\tau) d\tau\right) + \int_0^t kF\left(\frac{\hat{X}_k(t)}{k}\right) d\tau \quad \forall\ t \geq 0.\end{aligned}$$

The steady-state mean $k\rho$ of $\hat{X}_k$ increases to infinity as $k \to \infty$, and thus we must resort to some kind of normalization if we want to see a nondegenerate limit. A natural choice is to consider the processes $X_k = \hat{X}_k/k$, which have the same mean as the original process X. These processes satisfy the equations

$$\begin{aligned}X_k(t) = X_k(0) &+ \frac{1}{k}\left[Y_a(k\lambda t) - Y_d\left(\int_0^t k\mu X_k(t) t\right)\right] \\ &+ \int_0^t F\left(X_k(\tau)\right) d\tau \quad \forall\ t \geq 0.\end{aligned} \tag{1.4}$$

According to the strong law of large numbers for the Poisson process, provided in Appendix B, the second term in the right-hand side of the previous equation should vanish in the limit. Therefore, it is reasonable to expect that the processes X_k will

converge to a deterministic process solving the initial value problem $\dot{x} = F(x)$. More precisely, if we consider some initial condition $x_0 \geq 0$ and the solution x to the latter ODE, starting at x_0, then we have the following result.

Theorem 1.2.1. Assume that $X_k(0) \to x_0$ as $k \to \infty$, then

$$\sup_{t\in[0,T]} |X_k(t) - x(t)| \xrightarrow{\text{a.s.}} 0 \quad \forall\, T \geq 0.$$

We defer the proof of this strong law of large numbers until Chapter 2, where in fact we will prove a more general result. For now we provide the reader some simulations that advocate the veracity of our claim, these appear in Figure 1.3.

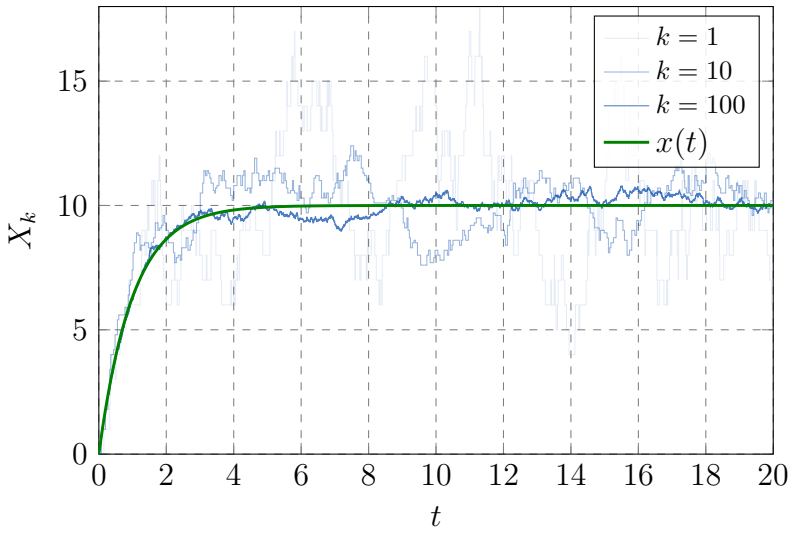


Figure 1.3: Paths of the processes X_k converging uniformly to x over a finite interval of time.

Note that the limit process x, usually referred to as fluid limit, solves an ODE that arises naturally from the drift of the chain:

$$\dot{x} = \lambda - \mu x. \tag{1.5}$$

As Figure 1.3 shows, this equation captures both the steady-state and transient behavior of the infinite-server queue, in the macroscopic scale.

If the number of requests is regarded as a real-valued variable, which makes sense in a large scale regime, then equation (1.5) represents a flow conservation law: the rate of change in the number of jobs equals the difference between the arrival rate of requests and their departure rate.

1.2.2 Central limit theorem

The normalization $X_k = \hat{X}_k/k$, that we adopted above, results in all the stochasticity vanishing as $k \to \infty$, leaving us with a deterministic limit. The rationale is that the standard deviation of the process $\hat{X}_k$, in the steady-state, is of order $\sqrt{k}$,

and thus normalizing by k causes stochastic fluctuations to disappear after taking the limit. The latter observation suggests that in order to recover some of the stochasticity, we could consider the processes $Z_k = \sqrt{k}(X_k - x)$, which represent the fluctuations of X_k around the fluid limit x, amplified by a factor of $\sqrt{k}$. Since the standard deviations of the processes X_k are of order $1/\sqrt{k}$, multiplication by $\sqrt{k}$ should compensate for the effect of our previous normalization.

If we write equation (1.5) in integral form, subtracting the resulting expression from equation (1.4), and multiplying by $\sqrt{k}$, we see that

$$\begin{aligned} Z_k(t) = Z_k(0) + \frac{1}{\sqrt{k}}\left[Y_a(k\lambda t) - Y_d\left(\int_0^t k\mu X_k(t)t\right)\right] \\ + \int_0^t \sqrt{k}\left[F\left(X_k(\tau)\right) - F(x(\tau))\right]d\tau \quad \forall\ t \geq 0. \end{aligned}$$

It is convenient to denote the middle term in the right-hand side by $U_k(t)$. Using this notation, and since F is an affine transformation, the above equation becomes

$$Z_k(t) = Z_k(0) + U_k(t) - \int_0^t \mu Z_k(\tau)d\tau \quad \forall\ t \geq 0. \tag{1.6}$$

It is possible to show that the process below is the solution to equation (1.6); details are given in Chapter 2 in a more general context.

$$Z_k(t) = Z_k(0) + U_k(t) - \int_0^t \mu e^{-\mu(t-\tau)}\left[Z_k(0) + U_k(\tau)\right]d\tau \quad \forall\ t \geq 0. \tag{1.7}$$

A right-continuous function with left-hand limits is called a càdlàg function, and the space of real-valued càdlàg functions that are defined on $[0, T]$ is denoted $D_{\mathbb{R}}[0, T]$ when it is endowed with the Skorohod topology; the main properties of this space are reviewed in Appendix A. The above equation (1.7) determines a mapping $\phi : D_{\mathbb{R}}[0, T] \longrightarrow D_{\mathbb{R}}[0, T]$ such that $Z_k = \phi(Z_k(0) + U_k)$. Furthermore, this map is continuous, as we will see in Chapter 2. Therefore, the continuous mapping theorem tells us that if $Z_k(0) + U_k$ had a limit in distribution, then the processes Z_k would have a limit in distribution as well.

As a matter of fact, if $Z_k(0) \to Z(0)$ as $k \to \infty$, for some constant $Z(0) \in \mathbb{R}$, then the processes U_k converge weakly in $D_{\mathbb{R}}[0, T]$. More precisely, if we let W_a and W_d be independent standard Wiener processes, then the limit in distribution of these processes is

$$U(t) = W_a(\lambda t) - W_d\left(\int_0^t \mu x(\tau)d\tau\right). \tag{1.8}$$

A detailed proof of this fact will be given in Chapter 3. For now we only tell the reader that the central limit theorem for the Poisson process, which can be found in Appendix B, will play an important role in the proof.

Returning to the processes Z_k, if we define $Z = \phi(Z(0) + U)$, then the condition $Z_k(0) \to Z(0)$ as $k \to \infty$ implies $Z_k \Rightarrow Z$ in $D_{\mathbb{R}}[0, T]$; where the process Z may also be regarded as the solution to the implicit integral equation

$$Z(t) = Z(0) + U(t) - \int_0^t \mu Z(\tau)d\tau \quad \forall\ t \geq 0. \tag{1.9}$$

As explained in Appendix A, convergence in $D_{\mathbb{R}}[0,T]$ for all $T \geq 0$ implies convergence in $D_{\mathbb{R}}[0,\infty)$. Since the time interval $[0,T]$ that we considered above is generic, this observation yields the following limit in distribution.

Theorem 1.2.2. Suppose that $Z_k(0) \to Z(0)$, for some constant $Z(0) \in \mathbb{R}$, as $k \to \infty$. Then $Z_k \Rightarrow Z$ in $D_{\mathbb{R}}[0,\infty)$ as $k \to \infty$, where Z solves equation (1.9). Moreover, the limit Z may also be regarded as the solution to the SDE

$$dZ_t = -\mu Z_t dt + \sqrt{\lambda + \mu x(t)} dW_t, \tag{1.10}$$

with initial condition $Z(0)$.

Note that the hypothesis of the last theorem implies $X_k(0) \to x_0$ as $k \to \infty$, hence Theorem 1.2.1 holds. Therefore, the processes Z_k indeed represent the small fluctuations of the processes X_k around the fluid limit x.

In order to justify the connection between equation (1.9) and the SDE (1.10), let W be a standard unidimensional Wiener process and let $f : [0,+\infty) \longrightarrow [0,+\infty)$ be a nonnegative and locally integrable function. In addition, consider the processes

$$W_1(t) = W\left(\int_0^t f(\tau)d\tau\right) \quad \text{and} \quad W_2(t) = \int_0^t \sqrt{f(\tau)} dW_\tau.$$

The latter Itô integral, seen as a function of its upper limit, defines a stochastic process. This process has a continuous martingale version and we are defining W_2 to be this version; we refer the reader to Appendix D. Equivalently, W_2 may be regarded as the solution to $dX_t = \sqrt{f(t)} dW_t$ when the initial condition is $X_0 = 0$. Note that the processes W_1 and W_2 are Gaussian, centered and have the same covariance, namely

$$\mathbb{E}\left[W_1(s)W_1(t)\right] = \int_0^s f(\tau)d\tau = \mathbb{E}\left[W_2(s)W_2(t)\right] \quad \forall\ s \leq t.$$

Therefore, they have the same finite-dimensional distributions.

Recall from equation (1.8) the definition of U as a sum of independent Wiener processes. Since the sum of two independent Gaussian random variables is also Gaussian, with variance the sum of the other two variances, then U has the same finite dimensional distributions as

$$W\left(\int_0^t \lambda + \mu x(\tau)d\tau\right) \sim \int_0^t \sqrt{\lambda + \mu x(\tau)} dW_\tau.$$

Here the term on the right has the same finite-dimensional distributions as the process on the left because of the observation at the end of the preceding paragraph.

Consequently, we see from equation (1.9) that Z has the same finite-dimensional distributions as the unique strong solution to equation (1.10). The results that are surveyed in Appendix D may be used to check that strong solutions to this SDE exist and are unique.

An important feature of equation (1.10) is that the drift coefficient is given by the derivative of F evaluated at the fluid limit x, specifically

$$dZ_t = F'(x(t))Z_t dt + \sqrt{\lambda + \mu x(t)} dW_t.$$

1.2.3 Steady-state estimates

We may now use the previous theorems to describe the steady-state behavior of the infinite-server queue in a large scale regime, that is with the traffic intensity ρ approaching infinity; the mean number of jobs equals ρ in steady-state, which justifies the terminology large scale. Before we provide this description, it is worth emphasizing that these results only allow to estimate the stationary distribution of the chain of Figure 1.2, whereas the actual distribution is Poisson of parameter ρ.

First, we observe that equation (1.5) has a single equilibrium point $x^* = \rho$, which is moreover a global attractor. This suggests that the processes X_k approach x^* as $t \to +\infty$, as it is shown in Figure 1.3. Therefore, if we want to understand the steady-state behavior of the infinite-server queue, it makes sense to set $x_0 = x^*$ in Theorem 1.2.1, so that the fluid limit of the infinite-server queue is the equilibrium solution $x \equiv x^*$ of the dynamics (1.5).

Under the above choice, the diffusion of Theorem 1.2.2 is an Ornstein-Uhlenbeck process. Indeed, if we set $x \equiv x^*$ in equation (1.10) then this SDE becomes

$$dZ_t = -\mu Z_t dt + \sqrt{2\lambda} dW_t.$$

The stationary distribution $Z(\infty)$ of the Ornstein-Uhlenbeck process is well-known, it is absolutely continuous and its density p may be derived by solving the Fokker-Planck equation associated to the latter SDE, under the condition that the solution must integrate one over the real line. This Fokker-Planck equation is

$$\lambda \frac{\partial^2 p}{\partial x^2} + \mu \frac{\partial (xp)}{\partial x} = 0$$

and its solution is the density of a centered Gaussian with variance ρ, namely

$$p(x) = \frac{1}{\sqrt{2\pi\rho}} e^{-\frac{x^2}{2\rho}}.$$

In order to interpret the above results, remember that Z_k represents the fluctuations of X_k around the fluid limit, in this case the equilibrium point x^*. Using the definitions of these processes we may write

$$\hat{X}_k = kx^* + \sqrt{k} Z_k,$$

where we recall that $\hat{X}_k$ is the number of jobs in an infinite-server queue with workload $k\rho$. Now Theorem 1.2.2 suggest the steady-state estimate

$$\hat{X}_k(\infty) \sim kx^* + \sqrt{k} Z(\infty).$$

Note that $kx^* = k\rho$ and $\sqrt{k} Z(\infty) \sim N(0, k\rho)$, therefore the last expression tells us that $\hat{X}_k(\infty)$ is approximately $N(k\rho, k\rho)$ when k is large enough. Since $k\rho$ is the traffic intensity that $\hat{X}_k$ faces, then we could incorporate the scaling in the estimate and say that the number of jobs in an infinite-server queue is approximately

$$X(\infty) \sim N(\rho, \rho)$$

in the steady-state and when the traffic intensity ρ is large enough.

As a final remark, we reconcile our previous Gaussian estimate with the fact that the steady-state distribution of the infinite-server queue is exactly Poisson of parameter ρ. To this end, recall that the stationary distribution of $\hat{X}_k$, that we computed from the chain, is Poisson of parameter $k\rho$, which is the distribution of the sum of k independent Poisson random variables of parameter ρ. Thus, by the central limit theorem for random variables, we have

$$Z_k(\infty) = \frac{\hat{X}_k(\infty) - k\rho}{\sqrt{k}} \Rightarrow N(0,\rho) \sim Z(\infty) \quad \text{in } \mathbb{R} \quad \text{as } k \to \infty.$$

1.3 Introducing further complexities

The infinite-server queue exhibits an ideal operation because customers do not experience any queueing delay and idle capacity does not exist. We would like to mimic this performance in practice, but the hurdle that we encounter is the delay in the execution of decisions within the cloud or data center infrastructure: it is not possible to spawn a server immediately, and neither can we get rid of idle servers right away. The lags in the creation and deletion of servers are random, and we will assume that they are exponential, to ensure that the model is still Markovian.

Let us imagine what the analog of the infinite-server queue would be like in the setting that we have described above. First, note that the algorithm behind the infinite-server queue can be described as follows.

- Immediately after a job departure, the infrastructure is asked to remove one server, and this action is executed right away.
- A new server is summoned whenever a job arrives and there are no idle servers; this new server is instantly created by the infrastructure.

In the presence of creation and deletion lags, servers cannot be dismissed or summoned right away, but we may consider the following alternative algorithm.

- A request is issued to the cloud or data center infrastructure, asking to shut down a server, immediately after each job departure; these requests are executed with an exponential delay of mean $1/c$ seconds.
- If a job arrives in the presence of idle servers, then the job is assigned to one of the idle servers and one of the shut down requests is withdrawn, if there are any of them pending.
- When a job arrives, and has to be queued, a new server is requested, but the infrastructure makes the server available only after an exponential time of mean $1/b$ seconds.
- If in the meanwhile one of the busy servers becomes idle, then the request is canceled and this idle server takes care of the queued job.

Note that the lags in the creation and deletion of servers, that we have introduced, prevent the number of servers and jobs from being equal. We are going to denote the number of servers in the system by M, whereas the number of jobs, either waiting in the queue or receiving service, will be called N; the stochastic process $X = (M, N)$ takes now values in the lattice $\mathbb{N}^2$.

In order to elucidate the Markovian model that describes the evolution of X over time, we first note that the number of pending server requests to the infrastructure is always equal to the number of queued jobs. Indeed, a new request is issued whenever a job is queued, and one request is removed when a job leaves the queue; either because the requested server appeared and took the job, or because an idle server took the job and one request was canceled. Similarly, the number of pending shut down requests is always equal to the number of idle servers: a request is made whenever a server finishes a job, and one request is canceled when an idle server takes a job; note that servers that appear in the system because they were summoned by the infrastructure become busy at once. Summing up, the number of pending server requests is $[N - M]^+$ and the number of pending shut down requests is $[M - N]^+$. Therefore, the dynamics of the queue that we have just described are given by the transitions diagram of Figure 1.4.

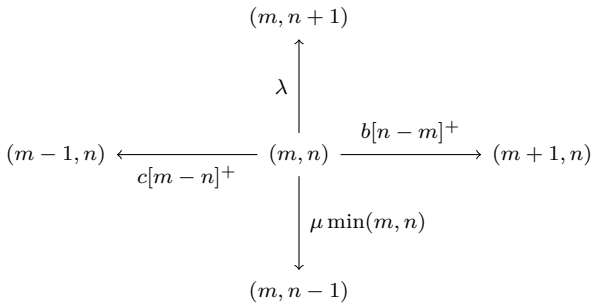


Figure 1.4: Markovian model of a system that attempts to emulate the infinite-server queue in the presence of non-negligible creation and deletion lags.

Computing the stationary distribution of this chain explicitly, from its balance equations, is at least a very difficult challenge, in contrast to the birth-death process that we studied in the previous section. Consequently, to understand the steady-state of the system, we must resort to the limit theorems that will be developed in the following chapters.

Even though we have not stated these theorems yet, we may extrapolate what we expect from the analysis of the last section's example. For instance, the drift of the chain is the field $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ that results from adding the intensities, regarded as vectors, that push away from a given state:

$$F(m, n) = \begin{bmatrix} b[n-m]^+ - c[m-n]^+ \\ \lambda - \mu \min(m, n) \end{bmatrix};$$

here we are letting the lower case m and n denote real numbers, representing the state of the system in the fluid scale. Comparing with Theorem 1.2.1 we expect that

the fluid limit $x = (m, n)$ of this chain will solve the initial value problem $\dot{x} = F(x)$ or, in coordinates, the ODE

$$\begin{aligned} \dot{m} &= b[n-m]^+ - c[m-n]^+, \\ \dot{n} &= \lambda - \mu \min(m, n). \end{aligned}$$

Note that the second of these equations is independent of the provisioning rule that we choose, or in other words the way in which we summon and dismiss servers. It is instead determined by the central queue scheme that we have adopted. The law of large numbers that we will prove in Chapter 2 yields the previous fluid limit under mild hypothesis on the intensities of Figure 1.4, the main of which is that F has to be locally Lipschitz, which is the case of the current chain.

We may also conjecture a central limit theorem for the chain of Figure 1.4. If we extrapolate the results of Subsection 1.2.2, a diffusion approximating this system's behavior should solve the SDE

$$dZ_t = A_t Z_t dt + B_t dW_t;$$

where W should be a bidimensional Wiener process, since we now have transitions in two possible directions, A_t should be the Jacobian matrix of F at the point $x(t)$ and B_t should be the matrix

$$B_t = \begin{bmatrix} b_{11}(t) & 0 \\ 0 & b_{22}(t) \end{bmatrix},$$

$$\text{with} \quad b_{11} = \sqrt{b[n-m]^+ + c[m-n]^+} \quad \text{and} \quad b_{22} = \sqrt{\lambda + \mu \min(m, n)}.$$

The definition of A_t only makes sense when $m(t) \neq n(t)$, because F is not differentiable along the diagonal. The central limit theorem that we will see in Chapter 2 requires the drift of the chain to be differentiable. However, in Chapter 3 we will extend this result to contemplate chains with non-differentiable drifts, which is the case of the current chain.

The mathematical background that we need in order to analyze chains like that of Figure 1.4 will be developed in the two following chapters. In Chapter 2 we will review the classical limit theorems due to Kurtz, for density dependent families of continuous time Markov chains. Afterwards, we will extend some of these results in Chapter 3; for instance, to contemplate chains with a non-differentiable drift. We will then return to the analysis of the queuing system that we have just described. Moreover, in Chapter 4 we will present alternative provisioning rules which aim at the elimination of queueing.

Chapter 2

Classical limit theorems

2.1 Density dependent families

In this section we specify the class of one parameter families of continuous time Markov chains that we will study throughout this chapter and the following. We will then prove a strong law of large numbers, analog of Theorem 1.2.1, and a central limit theorem, analog of Theorem 1.2.2, for these families of Markov chains. Before we do that, we introduce some convenient notation for the state-space and the transition rates of the Markov chains that will appear in the sequel.

Let $\mathbb{Z}^d \cup \{\Delta\}$ denote the one-point compactification of $\mathbb{Z}^d$. The Markov chains that we will consider take values on a subset of a the d-dimensional lattice, for starters given by the intersection of $\mathbb{Z}^d$ with some open set $E \subset \mathbb{R}^d$; the point Δ is reserved to denote the state of the chain after explosion. The possible jump directions will be given by a finite set $D \subset \mathbb{Z}^d$ and the transition rates will be determined by a family $\{\beta_l\}_{l\in D}$ of non-negative functions with domain E. We will assume that $x \in E \cap \mathbb{Z}^d$ and $\beta_l(x) > 0$ imply $x + l \in E \cap \mathbb{Z}^d$, this allows to define a continuous time Markov chain with state-space $E \cap \mathbb{Z}^d$ and intensities $q_{xy} = \beta_{y-x}(x)$.

Theorem 2.1.1. Consider a deterministic initial condition $X(0) \in E \cap \mathbb{Z}^d$ and let $\{\mathcal{N}_l\}_{l\in D}$ be an independent family of Poisson processes with unitary intensity, defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. There exists a unique stochastic process X such that

$$\begin{aligned} &X(t) = X(0) + \sum_{l\in D} l\mathcal{N}_l\left(\int_0^t \beta_l(X(\tau))d\tau\right) \quad \forall\ t \in [0,\zeta), \\ &X(t) = \Delta \quad \forall\ t \in [\zeta, +\infty) \quad \text{and} \\ &\zeta = \inf\left\{t \geq 0 : \lim_{s\to t^-} X(s) = \Delta\right\}; \end{aligned}$$

we are adopting the convention that the infimum of an empty set equals infinity. Furthermore, X is a continuous time Markov chain with state-space $E \cap \mathbb{Z}^d$, transition rates $q_{xy} = \beta_{y-x}(x)$, initial condition $X(0)$ and explosion time ζ.

This is the theorem that we used at the beginning of Subsection 1.2.1 to write the state of the infinite-server queue in terms of two independent Poisson processes. As mentioned there, the proof of this theorem is given in [8, Chapter 6.4] and its statement is also in line with one of the characterizations of continuous time Markov chains that are provided in [29, Chapter 2.6]. The importance of this theorem is that it will allow to construct sequences of continuous time Markov chains over the same probability space and, moreover, driven by the same family of Poisson processes. The first is clearly an essential condition for proving a strong law of large numbers.

Let us now introduce the notion of density dependent family. In order to do this, we will consider a sequence of spaces $S_k = E \cap k^{-1}\mathbb{Z}^d$ and we will assume that $x \in S_k$ and $\beta_l(x) > 0$ imply $x + k^{-1}l \in S_k$. As before, this hypothesis allows to define continuous time Markov chains with state-space S_k and transition rates q^k_{xy} that are proportional to $\beta_{k(y-x)}(x)$.

Definition 2.1.2. A density dependent family is a sequence of continuous time Markov chains X_k with state-space S_k and intensities $q^k_{xy} = k\beta_{k(y-x)}(x)$.

Note that as k increases the state-space S_k consists of a larger number of points which, furthermore, are closer to each other; for instance, if we compare S_k with S_1, neighboring states are k times closer. At the same time, as k grows the speed of transitions increases as well; indeed, if x lies both in S_1 and S_k, then transitions away from x occur k times faster in X_k than they occur in X_1. Informally speaking, the chain X_k jumps k times faster than X_1, but it covers a k times smaller distance every time it jumps. This yields the averaging phenomenon that we need to prove the law of large numbers of the following section.

Before we continue developing the theory, let us illustrate the above construction using the infinite-server queue as an example. The Markovian model of this queue, when the arrival rate is $k\lambda$, has as in Section 1.2 the following intensities.

$$q_{n,n+1} = k\lambda \quad \text{and} \quad q_{n,n-1} = \mu n = k\mu\frac{n}{k}.$$

Recall that n denotes the number of jobs in the system. In order to make the latter intensities fit into the framework of density dependent families we define the maps

$$\beta_1(x) = \lambda \quad \text{and} \quad \beta_{-1}(x) = \mu x,$$

and we see that the above rates may be rewritten in terms of these maps as

$$q_{n,n+1} = k\beta_1\left(\frac{n}{k}\right) \quad \text{and} \quad q_{n,n-1} = k\beta_{-1}\left(\frac{n}{k}\right);$$

the domain E of the maps β_l is discussed below. The intuition is that these intensities scale linearly with the parameter k and only depend on the "density" n/k; this name is inherited from epidemics models, where the latter fraction indeed represents a density, for instance the number of infected people among the whole population.

Finally, to define the chain X_k and make the infinite-server queue fit in the definition of density dependent families, we only need to introduce the change of variables $x = n/k$. This results in a Markov chain whose state x lies in $k^{-1}\mathbb{Z}$, and

has the following transition rates.

$$q_{x,x+k^{-1}} = k\beta_1(x) \quad \text{and} \quad q_{x,x-k^{-1}} = k\beta_{-1}(x).$$

In the present example, the natural choice of the set E would be $[0, +\infty)$, which is not an open set. Instead we may let E be any open set that contains $[0, +\infty)$; the form of the maps β_l ensures that the chain is confined to $[0, +\infty)$ if the initial condition lies inside of this interval. In general, we can always extend the natural domain of a density dependent family to an open set E, and this allows to adopt the convention that E is open. The latter is not essential, the proofs of the subsequent theorems may be carried out anyway, however this convention is convenient.

Returning to the general framework, we are going to let the maps β_l, that appear in Definition 2.1.2, depend on the scale parameter k, so that a broader class of sequences of continuous time Markov chains may fall into the category of density dependent families; this will be important in Chapter 4. Namely, differing with the notation that Kurtz uses, we are going to consider maps of the form

$$\beta_l^k = \gamma_l + \delta_l^k,$$

where β_l^k is still a non-negative map on E, such that $x \in S_k$ and $\beta_l^k(x) > 0$ imply $x + k^{-1}l \in S_k$. The terms δ_l^k are small perturbations, in the following sense.

Assumption 2.1.3. The next conditions hold inside of each compact set $K \subset E$.

$$\sup_{x\in K} |\delta_l^k(x)| < \infty \quad \forall\ l \in D,\ k \geq 1 \quad \text{and}$$
$$\lim_{k\to\infty} \sup_{x\in K} |\delta_l^k(x)| = 0 \quad \forall\ l \in D.$$

As mentioned above, Theorem 2.1.1 allows to construct the elements of a density dependent family over the same probability space. To do this we consider an independent family $\{\mathcal{N}_l\}_{l\in D}$ of Poisson processes with unitary intensity, defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and the maps

$$\hat{\beta}_l^k(u) = k\beta_l^k\left(\frac{u}{k}\right) \quad u \in kE \cap \mathbb{Z}^d.$$

Note that $u \in kE \cap \mathbb{Z}^d$ and $\hat{\beta}_l^k(u) > 0$ imply that $u + l \in kE \cap \mathbb{Z}^d$. Therefore, given some deterministic initial conditions $\hat{X}_k(0) \in kE \cap \mathbb{Z}^d$, by Theorem 2.1.1 it is possible to construct continuous time Markov chains $\hat{X}_k$ on $(\Omega, \mathcal{F}, \mathbb{P})$, with state-space $kE \cap \mathbb{Z}^d$ and transition rates $\hat{q}_{xy}^k = \hat{\beta}_{y-x}^k(x)$, such that

$$\hat{X}_k(t) = \hat{X}_k(0) + \sum_{l\in D} l\mathcal{N}_l\left(\int_0^t \hat{\beta}_l^k\left(\hat{X}_k(\tau)\right) d\tau\right)$$

holds for all t smaller than the explosion time $\hat{\zeta}_k$ of the chain. If we now consider the chains $X_k = \hat{X}_k/k$, then the sequence $\{X_k\}_{k\geq 1}$ is a density dependent family defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and its elements satisfy the equations

$$X_k(t) = X_k(0) + \sum_{l\in D} \frac{l}{k}\mathcal{N}_l\left(\int_0^t k\beta_l^k(X_k(\tau))d\tau\right) \quad \forall\ t \in [0, \zeta_k), \tag{2.1}$$

where $\zeta_k = \hat{\zeta}_k$ is the explosion time of X_k.

It is convenient to consider the centered Poisson processes $Y_l(t) = \mathcal{N}_l(t) - t$ rather than the the processes $\mathcal{N}_l$ themselves. Using the processes Y_l we may write

$$\sum_{l\in D} \frac{l}{k}\mathcal{N}_l\left(\int_0^t k\beta_l^k(X_k(\tau))d\tau\right) = \sum_{l\in D} \frac{l}{k}Y_l\left(\int_0^t k\beta_l^k(X_k(\tau))d\tau\right) + \int_0^t \sum_{l\in D} l\left[\gamma_l(X_k(\tau)) + \delta_l^k(X_k(\tau))\right]d\tau.$$

We will introduce the notation $\Sigma_k(t)$ to denote the first term on the right-hand side:

$$\Sigma_k(t) = \sum_{l\in D} \frac{l}{k}Y_l\left(\int_0^t k\beta_l^k(X_k(\tau))d\tau\right). \tag{2.2}$$

It is also convenient to introduce the following definition.

Definition 2.1.4. The drift and perturbing drifts of a density dependent family are, respectively, the vector fields $F, G_k : E \longrightarrow \mathbb{R}^d$ such that

$$F(x) = \sum_{l\in D} l\gamma_l(x) \quad \text{and} \quad G_k(x) = \sum_{l\in D} l\delta_l^k(x).$$

Using the above definition, we may rewrite equation (2.1) in terms of the drift and perturbing drift to yield

$$X_k(t) = X_k(0) + \Sigma_k(t) + \int_0^t F(X_k(\tau))d\tau + \int_0^t G_k(X_k(\tau))d\tau \quad \forall\, t \in [0, \zeta_k). \tag{2.3}$$

As in Section 1.2 this equation may be interpreted as an stochastically perturbed version of the initial value problem $\dot{x} = F(x)$.

2.1.1 An alternative approach

The purpose of this subsection is to comment on a different approach to the study of density dependent families; we will not adopt this approach, and thus the reader may skip this brief digression.

This alternative approach is based upon the following observation. If X is a Markov chain with infinitesimal generator Q, and f is a real-valued function defined on the the state-space of X, then

$$M_f(t) = f(X(t)) - f(X(0)) - \int_0^t Qf(X(\tau))d\tau$$

is a local martingale; we refer the reader to [32, Appendix B.3] and references therein.

Suppose now that $\{X_k\}_{k\geq 1}$ is a density dependent family whose elements are not necessarily defined over the same probability space. The infinitesimal generator Q_k of X_k acts on the identity e as follows.

$$Q_k e(x) = \sum_{l\in D} \left[e(x+l) - e(x)\right]\beta_l^k(x) = \sum_{l\in D} l\beta_l^k(x) = F(x) + G_k(x),$$

and applying the previous observation to each component of e, we see that

$$M_e^k(t) = X_k(t) - X_k(0) - \int_0^t F(X(\tau))d\tau - \int_0^t G_k(X(\tau))d\tau \tag{2.4}$$

is a vector local martingale. In particular, we have the following remark.

Remark 2.1.5. The process Σ_k of equation (2.3) is a local martingale. Particularly, if $\mathbb{E}\left(\sup_{s\in[0,t]} ||\Sigma_k(s)||\right) < \infty$ for all $t \geq 0$, then Σ_k is a martingale by [23, 8.a.4].

Equation (2.4) allows to prove a weak law of large numbers for density dependent families; we suggest the reader to look at [6]. However, the crucial advantage of the construction that leads to equation (2.3) is that the chains X_k are defined over the same probability space, and this is essential if we want to prove a strong law of large numbers. Another advantage of using equation (2.3) is that we may exploit the features of Poisson processes when we handle the local martingale term Σ_k; in fact, we will often do this rather than use generic martingale properties.

2.2 Strong law of large numbers

In this section our goal is to show that, under suitable hypothesis, there exists a set of probability one where the processes X_k converge uniformly over finite intervals of time to a deterministic process that solves the ODE $\dot{x} = F(x)$.

Lemma 2.2.1. Consider a bounded set $A \subset E$. Let $X = X_m$ for some fixed $m \geq 1$ and assume that $X(0) \in A$. With probability one X cannot take infinitely many jumps in A in finite time.

Proof. Let $\tau_i(\omega)$ be the time of the i-th jump of the path $X(\omega)$. As stated in [29, Theorem 2.8.4], conditional to $\mathcal{F}_n = \sigma(\{X_{\tau_i} : i = 0, \ldots, n\})$, the holding times $\{\tau_{i+1} - \tau_i : i = 0, \ldots, n\}$ are independent and exponential, with rates

$$\lambda_i = \sum_{l\in D} m\beta_l^m(X_{\tau_i}).$$

The boundedness of A implies that $A \cap S_m$ is finite, and thus these rates are uniformly bounded on the set $\Omega_A^n = \{\omega \in \Omega : X_{\tau_i}(\omega) \in A \ \forall\ i = 0, \ldots, n\}$ by the finite constant

$$\lambda = \sum_{l\in D} \sup_{x\in A\cap S_m} m\beta_l^m(x).$$

Define $\eta_i(\omega) = 1$ if $X_{\tau_i}(\omega) \in A$ and $\eta_i(\omega) = \infty$ if $X_{\tau_i}(\omega) \notin A$. Using the bound that we gave above we see that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\prod_{i=0}^n e^{-\eta_i(\tau_{i+1}-\tau_i)}\middle|\mathcal{F}_n\right]\right] &= \mathbb{E}\left[\mathbb{1}_{\Omega_A^n}\mathbb{E}\left[\prod_{i=0}^n e^{-(\tau_{i+1}-\tau_i)}\middle|\mathcal{F}_n\right]\right] \\ &= \mathbb{E}\left[\mathbb{1}_{\Omega_A^n}\prod_{i=0}^n \int_0^\infty e^{-t}\lambda_i e^{-\lambda_i t}dt\right] \leq \left(1+\frac{1}{\lambda}\right)^{-n};\end{aligned}$$

for the second equality we used the disintegration theorem [16, Theorem 5.4].

Letting $\mathcal{F}_\infty = \sigma(\{X_{\tau_i} : i \geq 0\})$ we get the following inequality by using dominated convergence twice; the first time applying [23, Theorem 4.c.2] to deal with the conditional expectations.

$$\begin{aligned}
\mathbb{E}\left[e^{-\sum_{i=0}^{\infty}\eta_i(\tau_{i+1}-\tau_i)}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{-\sum_{i=0}^{\infty}\eta_i(\tau_{i+1}-\tau_i)}\middle|\mathcal{F}_\infty\right]\right] \\
&= \mathbb{E}\left[\lim_{n\to\infty}\mathbb{E}\left[\prod_{i=0}^{n}e^{-\eta_i(\tau_{i+1}-\tau_i)}\middle|\mathcal{F}_n\right]\right] \\
&= \lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}\left[\prod_{i=0}^{n}e^{-\eta_i(\tau_{i+1}-\tau_i)}\middle|\mathcal{F}_n\right]\right] = 0.
\end{aligned}$$

This equation shows that, with probability one, X would need infinite time to jump infinitely many times inside of A. □

The last lemma is essentially a version of [29, Theorem 2.7.1] and, moreover, is implied in Theorem 2.1.1 in the definition of the time ζ. We will use this lemma jointly with the following, which provides a sort of induction principle.

Lemma 2.2.2. Let $f : [0, \eta] \longrightarrow E$ be a right continuous and piecewise constant function with finitely many jumps. Also, consider a proposition $P : E \longrightarrow \{0, 1\}$ and assume that:

1. $P(f(0)) = 1$.
2. $P(f(s)) = 1$ for all $s \in [0, t)$ implies $P(f(t)) = 1$.

Then $P(f(t)) = 1$ for all $t \in [0, \eta]$.

Proof. Suppose that there exists some $t_0 \in (0, \eta]$ such that $P(f(t_0)) = 0$. Since f is a right continuous and piecewise constant function with finitely many jumps, there exists $t = \min\{s \in [0, t_0] : P(f(s)) = 0\}$. This implies that $P(f(s)) = 1$ for all $s \in [0, t)$, and thus $P(f(t)) = 1$ contradicting the definition of t, and hence our initial assumption. We conclude that $P(f(t)) = 1$ for all $t \in [0, \eta]$. □

We will now continue under the following hypothesis.

Assumption 2.2.3. Suppose that F is locally Lipschitz and that the next condition holds inside of each compact set $K \subset E$.

$$\sup_{x\in K}|\gamma_l(x)| < \infty \quad \forall\, l \in D$$

Since F is now locally Lipschitz, we may consider the unique solution x to the ODE $\dot{x} = F(x)$, starting at some $x_0 \in E$ and defined on some interval $[0, T]$. We want to prove that the limit $X_k(0) \to x_0$ as $k \to \infty$ implies that

$$\sup_{t\in[0,T]}||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty.$$

This strong law of large numbers, which is the goal of this section, will be a straightforward consequence of the following result.

Lemma 2.2.4. Assume that $X_k(0) \to x_0$ as $k \to \infty$. Then there exist a compact neighborhood A of $\{x(t) : t \in [0,T]\}$, non-negative constants $\overline{\beta}_l$ and a null set $N \subset \Omega$ with the following property: for each $\omega \in N^c$ there exists $k_0(\omega)$ such that

$$\sup_{t\in[0,T]} ||X_k(\omega,t) - x(t)|| \leq \varepsilon_k(\omega)e^{MT} \quad \forall\ k \geq k_0(\omega);$$

where M is a Lipschitz constant for F inside the set A and

$$\varepsilon_k = ||X_k(0) - x_0|| + T\sup_{x\in A}||G_k(x)|| + \sum_{l\in D}||l||\sup_{t\in[0,T]}\frac{\left|Y_l(k\overline{\beta}_l t)\right|}{k}.$$

Proof. Fix some $\varepsilon > 0$ and note that since $\{x(t) : t \in [0,T]\}$ is compact, the neighborhood $A = \{y \in E : ||y - x(t)|| \leq \varepsilon \text{ for some } t \in [0,T]\}$ is compact as well, for a small enough ε. Therefore, it is possible to choose a uniform Lipschitz constant M inside of A for F.

By assumptions 2.1.3 and 2.2.3, the compactness of A implies that

$$\overline{\beta}_l = \sup_{k\geq 1, x\in A}\beta_l^k(x) < \infty \quad \forall\ l \in D.$$

Restating the hypothesis and using Assumption 2.1.3 again, together with the finiteness of the set of directions D, we also have

$$a_k = ||X_k(0) - x_0|| \to 0 \quad \text{and} \quad b_k = T\sup_{x\in A}||G_k(x)|| \to 0 \quad \text{as } k \to \infty.$$

These limits, together with the strong law of large numbers for the Poisson process, see Theorem B.1.1, and the finiteness of D, imply that

$$\varepsilon_k = a_k + b_k + \sum_{l\in D}||l||\sup_{t\in[0,T]}\frac{\left|Y_l(k\overline{\beta}_l t)\right|}{k} \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty.$$

Then, by Lemma 2.2.1 we may choose a null set $N \subset \Omega$ outside of which:

1. $\varepsilon_k \to 0$ as $k \to \infty$.

2. For each $k \geq 1$, such that $X_k(0) \in A$, the paths of X_k are right continuous, piecewise constant and take finitely many jumps in A in finite time.

Let $\delta > 0$ be such that $\delta + \delta e^{MT} \leq \varepsilon$. For each $\omega \in N^c$ we may choose $k_0(\omega)$ such that $k \geq k_0(\omega)$ implies that $\varepsilon_k(\omega) \leq \delta$, and also that $X_k(\omega)$ has jumps of length smaller than δ; the last condition may also be stated as

$$\max_{l\in D}||l|| \leq \delta k_0(\omega).$$

Fix some $\omega \in N^c$ and $k \geq k_0(\omega)$, we will show that $X_k(\omega,t) \in A$ for all $t \in [0,T]$. To this end, suppose the contrary, namely $\eta = \min\{t \geq 0 : X_k(\omega,t) \notin A\} \leq T$.

In order to arrive to a contradiction, fix $t \in (0,\eta]$ and assume that $X_k(\omega,s) \in A$ for all $s \in [0,t)$. Since $X_k(\omega)$ has finitely many jumps in $[0,\eta]$, then $\zeta_k(\omega) > \eta \geq t$,

and thus we may use equation (2.3). Moreover, for all $s \in [0,t)$ we have

$$\begin{aligned} &\int_0^s ||G_k(X_k(\omega,\tau))||\, d\tau \leq b_k \quad \text{and} \\ &||\Sigma_k(\omega,s)|| \leq \sup_{\tau\in[0,s]} ||\Sigma_k(\omega,\tau)|| \leq \varepsilon_k(\omega) - a_k - b_k, \end{aligned} \tag{2.5}$$

because $X_k(\omega,\tau) \in A$ for all $\tau \in [0,s]$; recall that Σ_k was defined in equation (2.2). These observations and equation (2.3) yield the following bound.

$$\begin{aligned} ||X_k(\omega,s) - x(s)|| &\leq a_k + b_k + ||\Sigma_k(\omega,s)|| + \int_0^s ||F(X_k(\omega,\tau)) - F(x(\tau))||\, d\tau \\ &\leq \varepsilon_k(\omega) + \int_0^s ||F(X_k(\omega,\tau)) - F(x(\tau))||\, d\tau \\ &\leq \delta + \int_0^s M\, ||X_k(\omega,\tau) - x(\tau)||\, d\tau \quad \forall\ s \in [0,t). \end{aligned}$$

An application of Gronwall's inequality now implies $||X_k(\omega,s) - x(s)|| \leq \delta e^{MT}$ for all $s \in [0,t)$. Recalling that $X_k(\omega)$ has jumps of length smaller than δ, this proves that $||X_k(\omega,t) - x(t)|| \leq \delta + \delta e^{MT} \leq \varepsilon$, and thus $X_k(\omega,t) \in A$. Then, $X_k(\omega,t) \in A$ for all $t \in [0,\eta]$ by Lemma 2.2.2; note that $X_k(0) \in A$ and since all jumps of $X_k(\omega)$, prior to time η, take place in A, then $X_k(\omega)$ has finitely many jumps in $[0,\eta]$.

The latter is a contradiction, because we had assumed that $X_k(\omega,\eta) \notin A$. Thus, we have shown that $X_k(\omega,t) \in A$ for all $t \in [0,T]$.

Finally, fix $\omega \in N^c$ and $k \geq k_0(\omega)$. Since $X_k(\omega,t) \in A$ for all $t \in [0,T]$, then the set of equations (2.5) holds, so we may write

$$||X_k(\omega,t) - x(t)|| \leq \varepsilon_k(\omega) + \int_0^t M\, ||X_k(\omega,\tau) - x(\tau)||\, d\tau \quad \forall\ t \in [0,T],$$

and the claim follows from another application of Gronwall's inequality. □

As it was announced, we now have the following theorem.

Theorem 2.2.5. Assume that $X_k(0) \to x_0$ as $k \to \infty$, then

$$\sup_{t\in[0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty.$$

Here the fluid limit x is the unique solution to

$$\dot{x} = F(x), \tag{2.6}$$

which we are assuming to be defined over the interval $[0,T]$.

Proof. Let M, N and ε_k be as in the statement of Lemma 2.2.4, then

$$\lim_{k\to\infty} \sup_{t\in[0,T]} ||X_k(\omega,t) - x(t)|| \leq \lim_{k\to\infty} \varepsilon_k(\omega) e^{MT} = 0 \quad \forall\ \omega \in N^c.$$

The limit on the right follows from the strong law of large numbers for the Poisson process; see Theorem B.1.1. Since N is a null set, this completes the proof. □

As a final remark, we note that it is possible to extend the last theorem in several

directions. First, one may show that the same result holds when the set of directions D is infinite; we refer the reader to [8, 22] for a proof. Second, it is possible to prove a law of large numbers for density dependent families whose drift is discontinuous, in this case the limit solves an inclusion differential equation and may not be unique; this is done in [10]. Finally, a weak law of large numbers holds in the more general context of pure jump Markov processes; the reader may find a proof in [20].

2.3 Central limit theorem

In order to motivate the developments of this section, we begin with the following observation: as we saw in Theorem 2.2.5, under mild assumptions, the error

$$\sup_{t\in[0,T]} ||X_k(t) - x(t)||$$

converges to zero as $k \to \infty$, and thus it is now natural to ask about the speed of convergence, or more precisely the order of magnitude of the latter error. Namely, we could ask how small α must be to ensure that the expression

$$\sup_{t\in[0,T]} k^\alpha \, ||X_k(t) - x(t)||$$

has still got a trivial limit. To give an answer to this question we will assume that

$$\lim_{k\to\infty} \sqrt{k} \sup_{x\in K} |\delta_l^k(x)| = 0 \quad \forall\ l \in D \tag{2.7}$$

holds inside each compact set $K \subset E$.

Theorem 2.3.1. Assume that $k^\alpha \, ||X_k(0) - x_0|| \to 0$ as $k \to \infty$, then

$$\sup_{t\in[0,T]} k^\alpha \, ||X_k(t) - x(t)|| \xrightarrow{\mathbb{P}} 0 \quad \text{as } k \to \infty \quad \forall\ \alpha \in [0, 1/2)$$

and the limit also holds almost surely for all $\alpha \in [0, 1/4)$.

Proof. Since $X_k(0) \to x_0$ as $k \to \infty$, Lemma 2.2.4 holds. Consider the definitions that we made in the statement of this lemma, and note that for $\omega \in N^c$ we have

$$\sup_{t\in[0,T]} k^\alpha \, ||X_k(\omega, t) - x(t)|| \le k^\alpha \varepsilon_k(\omega) e^{MT} \quad \forall\ k \ge k_0(\omega).$$

The hypothesis of this theorem and equation (2.7) imply that

$$a_k = k^\alpha \, ||X_k(0) - x_0|| + k^\alpha T \sup_{x\in A} ||G_k(x)|| \to 0 \quad \text{as } k \to \infty.$$

Therefore, it only remains to deal with the following term of $k^\alpha \varepsilon_k$.

$$b_k = k^\alpha \sum_{l\in D} ||l|| \sup_{x\in A} \frac{\left|Y_l(k\overline{\beta}_l t)\right|}{k}.$$

For $\alpha \in [0, 1/4)$ the above expression converges almost surely to zero as $k \to \infty$ by the strong law of large numbers for the Poisson process; see Theorem B.1.1.

For $\alpha \in [0, 1/2)$ fix some $\varepsilon > 0$ and let A_k be the set of the $\omega \in \Omega$ where

$$\sup_{t\in[0,T]} n^\alpha \, ||X_n(\omega, t) - x(t)|| \leq n^\alpha \varepsilon_n(\omega) e^{MT} \quad \forall \; n \geq k;$$

these sets increase to a set that contains N^c and hence has probability one. Also, Theorem B.1.1 implies that $b_k \xrightarrow{\mathbb{P}} 0$ as $k \to 0$, and thus there exists a sequence of sets $B_k \subset \Omega$ where $b_k e^{MT} < \varepsilon/2$ and such that $\mathbb{P}(B_k) \to 1$ as $k \to \infty$. Moreover, the inequality $a_k e^{MT} < \varepsilon/2$ holds on all Ω for large enough k, so we may write

$$\lim_{k\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} k^\alpha \, ||X_k(\omega, t) - x(t)|| < \varepsilon \right) \geq \lim_{k\to\infty} \mathbb{P}\left(A_k \cap B_k\right) = 1.$$

□

As a matter of fact, it turns out that $[0, 1/2)$ is the maximal interval with the following property: $\alpha \in [0, 1/2)$ and $k^\alpha \, ||X_k(0) - x_0|| \to 0$ as $k \to \infty$ imply that

$$\sup_{t\in[0,T]} k^\alpha \, ||X_k(t) - x(t)|| \xrightarrow{\mathbb{P}} 0 \quad \text{as } k \to \infty.$$

Indeed, as we will see below, for $\alpha = 1/2$ the process $k^\alpha(X_k - x)$ has a non-trivial limit in distribution, provided that $k^\alpha \left[X_k(0) - x_0\right]$ converges; this would answer the question that we posed at the beginning of this section completely.

Let us then address the case $\alpha = 1/2$. To this end let x be the solution to the initial value problem $\dot{x} = F(x)$, starting at x_0 and defined in $[0, T]$. Also let

$$Z_k = \sqrt{k}(X_k - x),$$

this process describes the fluctuations of X_k around the fluid limit x.

Theorem 2.3.2. Assume that the maps γ_l are locally Lipschitz, that the drift F is continuously differentiable, that assumptions 2.1.3 and 2.2.3 hold and also that equation (2.7) holds as well.

Suppose in addition that there exists some constant $Z(0) \in \mathbb{R}^d$ such that $Z_k(0) \to Z(0)$ as $k \to \infty$. Then $Z_k \Rightarrow Z$ in $D_{\mathbb{R}^d}[0, T]$ as $k \to \infty$, where Z is the continuous process such that

$$Z(t) = Z(0) + \sum_{l\in D} l W_l \left(\int_0^t \gamma_l(x(\tau)) d\tau \right) + \int_0^t F'(x(\tau)) Z(\tau) d\tau \quad \forall \; t \in [0, T]. \quad (2.8)$$

Here $\{W_l\}_{l\in D}$ is an independent family of standard unidimensional Wiener processes and F' is the Jacobian matrix of F. By the same arguments of Subsection 1.2.2 this process has the same finite-dimensional distributions as the solution to the SDE

$$dZ_t = F'(x(t)) Z_t dt + B_t dW_t$$

with initial condition $Z(0)$, where W is a d-dimensional Wiener process and

$$B_t = \sqrt{\sum_{l\in D} l l^T \gamma_l(x(t))}.$$

Here the square root is that of a positive semi-definite symmetric matrix.

A generalization of this theorem will be proved in Section 3.2, and therefore we will not provide a full proof here; nevertheless we will outline the proof that appears in [8, Chapter 11.2] and we will seize this opportunity to point out the differences with the arguments that will appear in Chapter 3.

To this purpose, we begin by observing that, if we subtract the integral form of the ODE $\dot{x} = F(x)$ from equation (2.3), and then multiply by $\sqrt{k}$, the result is

$$\begin{aligned} Z_k(t) = Z_k(0) + \sqrt{k}\Sigma_k(t) + \int_0^t \sqrt{k} G_k(X_k(\tau))d\tau \\ + \int_0^t \sqrt{k}\left[F(X_k(\tau)) - F(x(\tau))\right] d\tau \quad \forall\ t \in [0,T]. \end{aligned}$$

Furthermore, if we consider the series expansion of F around the point $x(t)$, to the first order, then we see that the last integrand equals

$$\sqrt{k}F'(x(\tau))\left[X_k(\tau) - x(\tau)\right] + \sqrt{k}R_\tau(X_k(\tau)) = F'(x(\tau))Z_k(\tau) + \sqrt{k}R_\tau(X_k(\tau)),$$

where R_τ is the first order remainder of the Taylor series around $x(\tau)$ and satisfies

$$\lim_{y \to x(\tau)} \frac{||R_\tau(y)||}{||y - x(\tau)||} = 0.$$

Using this observation we obtain the following equations for the processes Z_k.

$$Z_k(t) = Z_k(0) + U_k(t) + \delta_k(t) + \int_0^t F'(x(\tau))Z_k(\tau)d\tau \quad \forall\ t \in [0,T], \tag{2.9}$$

where U_k and δ_k are, respectively, the processes

$$\begin{aligned} U_k(t) = \sqrt{k}\Sigma_k(t) = \sum_{l \in D} \frac{l}{\sqrt{k}} Y_l\left(\int_0^t k\beta_l^k(X_k(\tau))d\tau\right) \quad \text{and} \\ \delta_k(t) = \int_0^t \sqrt{k}\left[R_\tau(X_k(\tau)) + G_k(X_k(\tau))\right] d\tau. \end{aligned}$$

Moreover, if we further let

$$U(t) = \sum_{l \in D} l W_l\left(\int_0^t \gamma_l(x(\tau))d\tau\right),$$

then we obtain a very similar equation for Z, indeed equation (2.8) becomes

$$Z(t) = Z(0) + U(t) + \int_0^t F'(x(\tau))Z(\tau)d\tau \quad \forall\ t \in [0,T]. \tag{2.10}$$

For each path outside a null set, equations (2.9) and (2.10) are of the form

$$\varphi(t) = f(t) + \int_0^t A(\tau)\varphi(\tau)d\tau \quad \forall\ t \in [0,T], \tag{2.11}$$

where $f \in D_{\mathbb{R}^d}[0,T]$ and $A(t)$ is a $d \times d$ matrix that varies continuously with respect to the parameter t. If we consider the fundamental matrix $\Phi(s,t)$ such that

$$\frac{\partial \Phi(s,t)}{\partial t} = A(t)\Phi(s,t) \quad \text{and} \quad \Phi(s,s) = \mathrm{Id} \quad \forall\ s,t \in [0,T],$$

then equation (2.11) may be solved explicitly, as it is proved below.

Lemma 2.3.3. There exists a unique function φ such that $\varphi - f$ is continuous and φ satisfies equation (2.11). Moreover, this function is

$$\varphi(t) = f(t) + \int_0^t \Phi(\tau, t) A(\tau) f(\tau) d\tau.$$

Proof. To begin, let us check that the latter expression solves equation (2.11). In order to do this we will compute, term by term, the integral

$$\int_0^t A(s)\varphi(s)ds = I_1 + I_2.$$

We are going to leave the first term I_1 unchanged,

$$I_1 = \int_0^t A(s) f(s) ds,$$

while for the integral I_2 of the second term we have

$$\begin{aligned} I_2 = \int_0^t A(s) \int_0^s \Phi(\tau, s) A(\tau) f(\tau) d\tau ds &= \int_0^t \int_0^s A(s)\Phi(\tau, s) A(\tau) f(\tau) d\tau ds \\ &= \int_0^t \int_0^s \frac{\partial \Phi(\tau, s)}{\partial s} A(\tau) f(\tau) d\tau ds \\ &= \int_0^t \frac{\partial}{\partial s} \left[\int_0^s \Phi(\tau, s) A(\tau) f(\tau) d\tau \right] ds \\ &\quad - \int_0^t \Phi(s, s) A(s) f(s) ds \\ &= \int_0^t \Phi(s, t) A(s) f(s) ds - \int_0^t A(s) f(s) ds. \end{aligned}$$

The fourth equality follows after applying Leibniz's rule; see Proposition E.1.1. Finally, adding $f(t) + I_1 + I_2$ we confirm that the solution that we proposed in the statement of the lemma indeed satisfies equation (2.11).

In order to check that this solution is unique, it is enough to observe that the difference between two solutions is continuous and satisfies the equation

$$\psi(t) = \int_0^t A(\tau)\psi(\tau)d\tau \quad \forall\ t \in [0, T],$$

which only has one continuous solution, this is $\psi \equiv 0$. □

This lemma allows to define a mapping $\phi : D_{\mathbb{R}^d}[0, T] \longrightarrow D_{\mathbb{R}^d}[0, T]$ carrying each $f \in D_{\mathbb{R}^d}[0, T]$ to the unique $\phi_f \in D_{\mathbb{R}^d}[0, T]$ such that $\phi_f - f$ is continuous and

$$\phi_f(t) = f(t) + \int_0^t F'(x(\tau))\phi_f(\tau)d\tau \quad \forall\ t \in [0, T].$$

An explicit construction of this function will no longer be possible in Section 3.4, where we will consider density dependent families with a non-differentiable drift.

Note that in equation (2.9) the integrand $F'(x(\tau))Z_k(\tau)$ is a càdlàg function, and therefore the corresponding integral, seen as a function of the upper limit, defines a continuous map. Equivalently, the function $Z_k - Z_k(0) - U_k - \delta_k$ is continuous, and hence $Z_k = \phi(Z_k(0) + U_k + \delta_k)$. As a result, in order to prove Theorem 2.3.2

it suffices to show that ϕ is a continuous map and $Z_k(0) + U_k + \delta_k \Rightarrow Z(0) + U$ in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$. Indeed, if we could prove that, then the continuous mapping theorem would yield the weak convergence of the processes Z_k to the continuous process $Z = \phi(Z(0) + U)$; here Z satisfies equation (2.10) by the definition of ϕ.

By Proposition A.1.8, in order to prove that $Z_k(0) + U_k + \delta_k \Rightarrow Z(0) + U$ in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$, it is enough to show that $U_k \Rightarrow U$ in $D_{\mathbb{R}^d}[0,T]$ and

$$\sup_{t\in[0,T]} ||\delta_k(t)|| \xrightarrow{\mathbb{P}} 0 \quad \text{as } k \to \infty;$$

note that the hypothesis of Theorem 2.3.2 already says that $Z_k(0) \to Z(0)$.

A proof of the limit $U_k \Rightarrow U$, in the context of pure jump Markov processes and under very general conditions, may be found in [21]. However, the proof that we are going to give in Chapter 3 is different and it is based on the central limit theorem for the Poisson process. There we will also deal with the processes δ_k, which in the setting of Chapter 3 will converge in probability to a deterministic process, not necessarily zero; here the limit is zero as a consequence of equation (2.7).

Thus, we will defer the proof of $Z_k(0) + U_k + \delta_k \Rightarrow Z(0) + U$ until Chapter 3. If we assume this fact for now, then in order to complete the proof of Theorem 2.3.2 it only remains to be shown that the mapping ϕ is continuous, and this is done below.

Lemma 2.3.4. The map $\phi : D_{\mathbb{R}^d}[0,T] \longrightarrow D_{\mathbb{R}^d}[0,T]$ is continuous.

Proof. Fix some $f \in D_{\mathbb{R}^d}[0,T]$, by Lemma 2.3.3 we know that

$$\phi_f(t) = f(t) + \int_0^t \Gamma(\tau, t) f(\tau) d\tau,$$

where $\Gamma(s,t) = \Phi(s,t)F'(x(s))$ is a continuous function.

To prove that ϕ is continuous at f, we will work with one of the metrics that generate the Skorohod topology, namely d_0. Given some $g \in D_{\mathbb{R}^d}[0,T]$, the distance $d_0(f,g)$ is defined as the infimum of those $\delta > 0$ for which there exists an increasing and continuous bijection $\lambda : [0,T] \longrightarrow [0,T]$ with the following properties.

$$\sup_{s,t\in[0,T]} \left| \log\left(\frac{\lambda(t) - \lambda(s)}{t - s} \right) \right| \leq \delta \quad \text{and} \quad \sup_{t\in[0,T]} ||f(\lambda(t)) - g(t)|| \leq \delta;$$

further details are provided in Appendix A. To show that $d_0(f,g) \to 0$ implies $d_0(\phi_f, \phi_g) \to 0$, we first note that given $\delta \geq d_0(f,g)$ there exists some λ_δ with the above characteristics; we will drop the subscript δ to simplify the notation.

The first property implies that $|\lambda(t) - \lambda(s)| \leq e^\delta |t - s|$ for all $s, t \in [0,T]$, which means that λ is Lipschitz and in particular absolutely continuous. Hence, its derivative exists almost everywhere on $[0,T]$, and at the points where it exists we have $|\lambda'(t) - 1| \leq e^\delta - 1$. Using the absolute continuity of λ we may write

$$\begin{aligned}
\phi_f(\lambda(t)) - \phi_g(t) &= f(\lambda(t)) - g(t) + \int_0^{\lambda(t)} \Gamma(\tau, \lambda(t)) f(\tau) d\tau - \int_0^t \Gamma(s,t) g(s) ds \\
&= f(\lambda(t)) - g(t) + \int_0^t \Gamma(\lambda(s), \lambda(t)) f(\lambda(s)) \lambda'(s) ds - \int_0^t \Gamma(s,t) g(s) ds.
\end{aligned}$$

Operating on the right-hand side we see that

$$\begin{aligned}\phi_f(\lambda(t)) - \phi_g(t) = f(\lambda(t)) - g(t) &+ \int_0^t \Gamma(\lambda(s),\lambda(t)) f(\lambda(s)) \left[\lambda'(s) - 1\right] ds \\ &+ \int_0^t \left[\Gamma(\lambda(s),\lambda(t)) - \Gamma(s,t)\right] f(\lambda(s)) ds \\ &+ \int_0^t \Gamma(s,t) \left[f(\lambda(s)) - g(s)\right] ds,\end{aligned}$$

therefore we have

$$\begin{aligned}\sup_{t\in[0,T]} ||\phi_f(\lambda(t)) - \phi_g(t)|| \leq &\sup_{t\in[0,T]} ||f(\lambda(t)) - g(t)|| \\ &+ \sup_{s,t\in[0,T]} T\,||\Gamma(\lambda(s),\lambda(t))||\,||f(\lambda(s))||\,(e^\delta - 1) \\ &+ \sup_{s,t\in[0,T]} T\,||\Gamma(\lambda(s),\lambda(t)) - \Gamma(s,t)||\,||f(\lambda(s))|| \\ &+ \sup_{s,t\in[0,T]} T\,||\Gamma(s,t)||\,||f(\lambda(s)) - g(s)||\,,\end{aligned}$$

Since Γ is continuous we know that Γ is bounded in $[0,T]^2$. Also, the fact that f is càdlàg implies that f is bounded in $[0,T]$ as well, thus the second term on the right hand side converges to zero as $\delta \to 0$.

The third term also converges to zero as $\delta \to 0$. To prove this note that

$$|\lambda(t) - t| = t\left|\frac{\lambda(t) - \lambda(0)}{t - 0} - 1\right| \leq t(e^\delta - 1) \leq T(e^\delta - 1) \quad \forall\, t \in [0,T].$$

Using the continuity of Γ and the compactness of $[0,T]^2$ we may then see that $||\Gamma(\lambda(s),\lambda(t)) - \Gamma(s,t)|| \to 0$ as $\delta \to 0$.

Finally, the first and fourth terms converge to zero as $\delta \to 0$ as well, because

$$\sup_{t\in[0,T]} ||f(\lambda(t)) - g(t)|| \leq \delta.$$

This proves that ϕ is continuous at f. □

2.3.1 Characterization of the limit

In the central limit theorem that we stated above, the limit

$$Z(t) = Z(0) + U(t) + \int_0^t \Phi(\tau,t) F'(x(\tau)) \left[Z(0) + U(\tau)\right] d\tau \tag{2.12}$$

is a time inhomogeneous Gaussian process. Below we prove this fact and we then compute the mean and covariance of Z.

In order to prove this, we begin by noticing that the process U, that appears in equation (2.10), is a time inhomogeneous Wiener process. Indeed, we recall that

$$U(t) = \sum_{l\in D} l W_l\left(\int_0^t \gamma_l(x(\tau)) d\tau\right).$$

It is clear that we only need to show that the sum of the last two terms of equation (2.12) is a Gaussian process, as a function of t. To this end, consider a sequence of partitions $0 = s_0^n < \cdots < s_{r_n}^n = T$ such that

$$\lim_{n\to\infty} \max_{1\le i\le r_n} s_i^n - s_{i-1}^n = 0.$$

Fix some $t \in [0,T]$, let $t_i^n = t \wedge s_i^n$ and define $\Delta_i^n = t_i^n - t_{i-1}^n$ to be the length of the interval $[s_{i-1}^n, s_i^n] \cap [0,t]$. Then by the continuity of U we have

$$U(t) + \sum_{i=1}^{r_n} \Phi(t_i^n, t) F'(x(t_i^n)) U(t_i^n) \Delta_i^n \xrightarrow{\text{a.s.}} U(t) + \int_0^t \Phi(\tau, t) F'(x(\tau)) U(\tau) d\tau$$

as $n \to \infty$. Since the summations on the left define Gaussian processes, then their limit is also of this kind, and thus Z is a Gaussian process.

Before we compute the mean and covariance of Z, we recall that this process has the same law as the solution to the SDE

$$dZ_t = A_t Z_t dt + B_t dW_t, \tag{2.13}$$

where A_t is the Jacobian matrix of F at $x(t)$, W_t is a d-dimensional Wiener process, with independent coordinates, and B_t is the $d \times d$ matrix

$$B_t = \sqrt{\sum_{l\in D} l l^T \gamma_l(x(t))}.$$

If we take expectations on both sides of the integral version of equation (2.13), then the Itô integral vanishes. Therefore, the mean of Z satisfies

$$\mu_t = Z_0 + \int_0^t A_\tau \mu_\tau d\tau,$$

and using the fundamental matrix Φ to solve this equation, we get $\mu(t) = \Phi(0,t) Z(0)$.

Proposition 2.3.5. The covariance of Z is

$$\Sigma(s,t) = \mathbb{E}\left[(Z_s - \mu_s)(Z_t - \mu_t)^T \right] = \int_0^s \Phi(\tau, s) B_\tau B_\tau^T \Phi(\tau, t)^T d\tau$$

for all $0 \le s \le t \le T$.

Proof. The centered process $Z - \mu$ solves equation (2.13) when the initial condition is zero. Thus, we may assume that $Z_0 = 0$, or equivalently that Z is centered.

By the Itô formula, the matrix process $X_t = Z_t Z_t^T$ satisfies the SDE

$$\begin{aligned} dX_t &= (dZ_t) Z_t^T + Z_t (dZ_t)^T + B_t B_t^T dt \\ &= (A_t X_t + X_t A_t^T + B_t B_t^T) dt + (B_t dW_t) Z_t^T + Z_t (B_t dW_t)^T. \end{aligned}$$

If we take expectations in both sides of the above equation, then the integrals with respect to Wiener processes vanish. Therefore, letting $x_t = \mathbb{E}[X_t]$ and noting that $x_0 = 0$, because $Z_0 = 0$, we see that

$$x_t = \int_0^t A_\tau x_\tau + x_\tau A_\tau^T + B_\tau B_\tau^T d\tau.$$

We are going to check that

$$x_t = \int_0^t G(\tau, t) d\tau \quad \text{where} \quad G(\tau, t) = \Phi(\tau, t) B_\tau B_\tau^T \Phi(\tau, t)^T.$$

To this purpose, we use Leibniz's rule to compute

$$\begin{aligned} A_\tau \int_0^\tau G(\sigma, \tau) d\sigma + \int_0^\tau G(\sigma, \tau) d\sigma A_\tau^T = \int_0^\tau \frac{\partial G(\sigma, \tau)}{\partial \tau} d\sigma = \frac{\partial}{\partial \tau} \int_0^\tau G(\sigma, \tau) d\sigma - G(\tau, \tau) \\ = \frac{\partial}{\partial \tau} \int_0^\tau G(\sigma, \tau) d\sigma - B_\tau B_\tau^T. \end{aligned}$$

If we now integrate the first term in the right-hand side, then we have

$$\int_0^t \frac{\partial}{\partial \tau} \left[\int_0^\tau G(\sigma, \tau) d\sigma \right] d\tau = \int_0^t G(\tau, t) d\tau;$$

this proves that $\Sigma(t, t)$ is as we claimed.

Now fix some $s \in [0, T]$ and consider the matrix process $Y_t = Z_s Z_t^T$ defined in the interval $[s, T]$. Note that this process satisfies

$$dY_t = Z_s Z_t^T A_t^T dt + Z_s (B_t dW_t)^T = Y_t A_t^T dt + Z_s (B_t dW_t)^T.$$

When we take the expectation of the right-hand side, the last term vanishes because it is a martingale. Hence, if we let $y_t = \mathbb{E}[Y_t]$, then we have

$$y_t = \Sigma(s, s) + \int_s^t y_\tau A_\tau^T d\tau.$$

Since Z is continuous, we know that y is continuous as well. Recall that A_t is also continuous, thus the fact that y solves this equation implies that y is differentiable. As a result, y solves the ODE $\dot{y}_t = y_t A_t^T$, starting at $y_s = \Sigma(s, s)$.

The solution to this equation is $y_t = \Sigma(s, s) \Phi(s, t)^T$, or equivalently

$$\begin{aligned} y_t &= \int_0^s \Phi(\tau, s) B_\tau B_\tau^T \Phi(\tau, s)^T d\tau \Phi(s, t)^T \\ &= \int_0^s \Phi(\tau, s) B_\tau B_\tau^T \Phi(\tau, s)^T \Phi(s, t)^T d\tau = \int_0^s \Phi(\tau, s) B_\tau B_\tau^T \Phi(\tau, t)^T d\tau. \end{aligned}$$

□

2.4 Affine and stable drifts

In this section we assume that the drift F is an affine vector field with an stable Jacobian matrix A, by this we mean that A has eigenvalues with strictly negative real parts. We will suppose in addition that $\delta_l^k \equiv 0$ for all $l \in D$ and $k \geq 1$ and we will introduce the technical hypothesis that the following function is affine.

$$x \longmapsto \sum_{l \in D} \gamma_l(x)$$

Below we first prove that, in this case, the chains X_k have an almost surely infinite explosion time; here we use our technical hypothesis. Afterwards, we observe that

$\mathbb{E}[X_k]$ solves the fluid equation $\dot{x} = F(x)$, the ODE that governs the fluid limit of the density dependent family. Then, we study the limit of Theorem 2.3.2 when we take an equilibrium point of $\dot{x} = F(x)$ as nominal solution. In particular, we see that the matrix B_t of equation (2.13) is constant and we use this to characterize the steady-state of the limit, showing that its covariances solves a Lyapunov equation.

Proposition 2.4.1. Let $X = X_m$ for some fixed $m \geq 1$, then X is non-explosive.

Proof. As in the proof of Lemma 2.2.1 let $\tau_i(\omega)$ be the time of the i-th jump of the path $X(\omega)$. Remember that conditional to $\mathcal{F}_n = \sigma\left(\{X_{\tau_i} : i = 0, \ldots, n\}\right)$, the holding times $\{\tau_{i+1} - \tau_i : i = 0, \ldots, n\}$ are independent and exponential with rates

$$\lambda_i = \sum_{l \in D} m\gamma_l(X_{\tau_i}).$$

Furthermore, since this is an affine map, as a function of X_{τ_i}, then there exist a constant $\alpha \in \mathbb{R}$ and a linear transformation $S : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $\lambda_i = \alpha + S(X_{\tau_i})$. Therefore, these rates are bounded by

$$\lambda_i \leq |\alpha| + ||S||\,||X_{\tau_i}|| \leq |\alpha| + ||S||\left(||X(0)|| + i \max_{l \in D} ||l||\right) = a + bi,$$

where a and b are non-negative constants. This implies that

$$\prod_{i=0}^{\infty}\left(1 + \frac{1}{\lambda_i}\right) > \sum_{i=1}^{\infty}\frac{1}{\lambda_i} \geq \sum_{i=0}^{\infty}\frac{1}{a+bi} = +\infty.$$

As a result, using the disintegration theorem [16, Theorem 5.4] as in the proof of Lemma 2.2.1, we see that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}\left[\prod_{i=0}^{n} e^{-(\tau_{i+1}-\tau_i)}\middle|\mathcal{F}_n\right]\right] &= \lim_{n\to\infty}\mathbb{E}\left[\prod_{i=0}^{n}\int_0^{\infty} e^{-t}\lambda_i e^{-\lambda_i t}dt\right] \\ &= \lim_{n\to\infty}\mathbb{E}\left[\prod_{i=0}^{n}\left(1+\frac{1}{\lambda_i}\right)^{-1}\right] = 0.\end{aligned}$$

Hence, the same computation that we used at the end of Lemma 2.2.1 shows that

$$\mathbb{E}\left[e^{-\sum_{i=0}^{\infty}(\tau_{i+1}-\tau_i)}\right] = \lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}\left[\prod_{i=0}^{n} e^{-(\tau_{i+1}-\tau_i)}\middle|\mathcal{F}_n\right]\right] = 0.$$

Thus, X needs infinite time to accomplish infinitely many jumps, almost surely. □

In the setting of this section, equation (2.3) simplifies to

$$X_k(t) = X_k(0) + \Sigma_k(t) + \int_0^t F(X_k(\tau))d\tau \quad \forall\ t \geq 0 \quad \text{a.s.} \tag{2.14}$$

Because $G_k \equiv 0$ and the explosion time of X_k is almost surely infinite for all $k \geq 0$.

Taking expectations on both sides of equation (2.14) we have

$$\mathbb{E}[X_k(t)] = X_k(0) + \int_0^t F(\mathbb{E}[X_k(\tau)])d\tau \quad \forall\ t \geq 0,$$

because under the hypothesis of this section one can check that Σ_k is a martingale by Remark 2.1.5. The mean of X_k thus solves $\dot{x} = F(x)$; since F is affine, solutions to this ODE are defined for all times $t \geq 0$.

Recall that the fluid limit solves this ODE as well, and since solutions are defined for all $t \geq 0$, we can say that the limit $X_k(0) \to x_0$ as $n \to \infty$ implies

$$\sup_{t\in[0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty \quad \forall\, T \geq 0,$$

where x is the solution to $\dot{x} = F(x)$ that starts at x_0.

The fact that A is stable implies that $\dot{x} = F(x)$ has a global attractor x^*. Moreover, it is always possible to find a positive definite symmetric matrix P such that $A^T P + PA$ is negative definite. This allows to construct a quadratic Lyapunov function $V(y) = (y - x^*)^T P(y - x^*)$ for the fluid dynamics.

Note that x^* has the property that $X_k(0) \to x^*$ as $k \to \infty$ implies

$$\sup_{t\in[0,T]} ||X_k(t) - x^*|| \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty \quad \forall\, T \geq 0.$$

2.4.1 The Lyapunov equation

The diffusion that appears in the claim of Theorem 2.3.2, for a generic solution x to the fluid dynamics, is given in this case by the SDE

$$Z_t = AZ_t dt + B_t dW_t, \tag{2.15}$$

where W is a d-dimensional Wiener process and B is as in Subsection 2.3.1. Since solutions to $\dot{x} = F(x)$ are defined for all $t \geq 0$, Theorem 2.3.2 yields weak convergence in $D_{\mathbb{R}^d}[0,T]$ for all $T \geq 0$. This implies that the limit takes place in $D_{\mathbb{R}^d}[0,\infty)$ as well, by Theorem A.3.6.

Note that the fundamental matrix of $\dot{x} = F(x)$ is $\Phi(s,t) = e^{A(t-s)}$, therefore the mean and covariance of Z are given by

$$\mu(t) = e^{At} Z(0) \quad \text{and} \quad \Sigma(s,t) = \int_0^s e^{A(s-\tau)} B_\tau B_\tau^T e^{A^T(t-\tau)} d\tau \quad \forall\, 0 \leq s < t.$$

In the special case where the nominal solution to the fluid dynamics is an equilibrium point $x \equiv x^*$, the dispersion coefficient B turns out to be constant in time:

$$B = \sqrt{\sum_{l\in D} ll^T \gamma_l(x^*)}.$$

In this case $\Sigma(t) = \Sigma(t,t)$ solves $\dot{\Sigma} = A\Sigma + \Sigma A^T + BB^T$. Since A is stable, $\Sigma(t)$ has a limit Σ_∞ when $t \to +\infty$. Moreover, this implies that $\dot{\Sigma}(t)$ has a limit as well, because $\dot{\Sigma}(t) = A\Sigma(t) + \Sigma(t)A^T + BB^T$. This limit is necessarily zero, otherwise $\Sigma(t)$ would not converge as $t \to +\infty$. Hence, the matrix Σ_∞ solves

$$A\Sigma_\infty + \Sigma_\infty A^T + BB^T = 0. \tag{2.16}$$

Assuming that the invariant measure of equation (2.15) exists, this Lyapunov equa-

tion provides an easy way to compute its covariance matrix. Below we give a condition that ensures the existence of this measure. Moreover, we prove that under this condition the stationary distribution is a centered Gaussian whose covariance solves the above Lyapunov equation.

Suppose that the diffusion coefficient BB^T is non-singular; this condition may be weakened as it is explained in Appendix D. We may then use the Foster-Lyapunov criteria of the same appendix to prove that a unique invariant measure exists and is exponentially ergodic. To this end, consider the quadratic Lyapunov function $V(y) = (y - x^*)^T P(y - x^*)$ that we mentioned above, when we discussed the global asymptotic stability of the fluid dynamics.

To use the above mentioned Foster-Lyapunov criteria, we need to work with the second order differential operator L that characterizes the SDE (2.15), specifically

$$Lf(y) = \nabla f(y)Ay + \frac{1}{2}tr\left[BH_f(y)B^T\right] \quad \forall\ f \in C^2(\mathbb{R}^d). \tag{2.17}$$

The reader may see appendices C and D for further details.

Since the second order derivatives of V are constant, the second term of (2.17) is equal to some constant κ when we replace f by V. Hence, we see that

$$\begin{aligned} LV(y) &= 2(y - x^*)^T PAy + \kappa \\ &= 2(y - x^*)^T PA(y - x^*) + 2(y - x^*)^T PAx^* + \kappa \\ &= (y - x^*)^T (A^T P + PA)(y - x^*) + 2(y - x^*)^T PAx^* + \kappa, \end{aligned}$$

Recall that $A^T P + PA$ is negative definite, whereas P is positive definite. Then, there exists a constant $c > 0$ such that

$$(y - x^*)^T (A^T P + PA)(y - x^*) < -c(y - x^*)^T P(y - x^*) = -cV(y) \quad \forall\ y \in \mathbb{R}^d.$$

The left-hand side is quadratic in the coefficients of y, while the other terms in $LV(y)$ are at most linear in these coefficients. This implies that

$$\liminf_{y\to\infty} \frac{LV(y)}{-cV(y)} = \liminf_{y\to\infty} \frac{(y - x^*)^T (A^T P + PA)(y - x^*)}{-cV(y)} > 1.$$

Therefore, there exists $r > 0$ such that $LV(y) \leq -cV(y)$ whenever $||y|| > r$. If we define $d = \max\{LV(y) : ||y|| \leq r\}$, then $LV(y) \leq -cV(y) + d$ for all $y \in \mathbb{R}^d$. Hence, by the Foster-Lyapunov criteria of Subsection D.3.1, we know that equation (2.15) admits a exponentially ergodic invariant measure.

Let Z_∞ be distributed according to the invariant measure of equation (2.15) and consider any solution Z to this SDE, starting at some point $Z_0 \in \mathbb{R}^d$. The exponential ergodicity of equation (2.15) implies that $Z_t \Rightarrow Z_\infty$ in $\mathbb{R}^d$ as $t \to +\infty$; see Section D.3.1. In particular we have

$$\mathbb{E}\left[e^{i\langle y, Z_\infty\rangle}\right] = \lim_{t\to+\infty} \mathbb{E}\left[e^{i\langle y, Z_t\rangle}\right] = \lim_{t\to+\infty} e^{-\frac{1}{2}y^T\Sigma(t)y + i\mu(t)^T y} = e^{-\frac{1}{2}y^T\Sigma_\infty y}.$$

Thus, Z_∞ is a centered Gaussian whose covariance solves equation (2.16).

Chapter 3

Extensions of the central limit theorem

3.1 A motivating example

In order to illustrate the problems that this chapter addresses, we begin with a classical example: the heavy traffic analysis of the many-server queue in the Halfin-Whitt regime; the reader may find the original work in [14].

The many-server queue models a data center with fixed capacity: it is a queueing system consisting of a centralized queue and a finite number of servers. Under the exponential assumptions of Section 1.1, if we let m be the number of servers, then the number of jobs n evolves according to the birth-death process of Figure 3.1. We are letting λ_m be the arrival rate of jobs and we are denoting by μ the service rate of each server; note that at any given time the number of active servers is $\min(m, n)$.

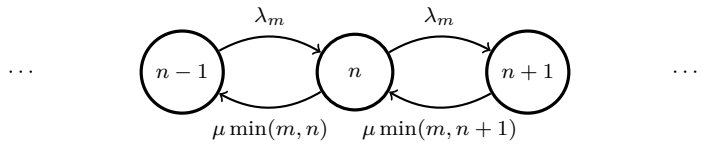


Figure 3.1: Birth-death process describing the number of jobs in a many-server queue.

In order to understand the behavior of a large scale many-server queue, we may let λ_m and m approach infinity; the condition $\lambda_m/m\mu < 1$ must be enforced if we want the chain to be stable, and thus we need to scale the arrival rate and the number of servers simultaneously. Halfin and Whitt proposed a scaling of the form

$$\lambda_m = m\mu - O\left(\sqrt{m}\right).$$

The result is a sequence of many-server queues approaching heavy traffic: they work increasingly closer to the border of their capacity as the number of servers grows, because $\lambda_m/m\mu \to 1$. Moreover, we will see that after taking the limit the

steady-state probability of finding idle capacity is positive but strictly smaller than one. This is the situation in real-life data centers, which are designed for a certain tradeoff between user perceived performance and operation costs; this feature is what works prior to [14] failed to capture.

Remark 3.1.1. Let $\rho_m = \lambda_m/\mu$ be the traffic intensity, in the Halfin-Whitt scaling $\rho_m/m \to 1$ as $m \to \infty$. Thus, $m = \rho_m + O(\sqrt{\rho_m})$ which yields the square root staffing rule: estimate the traffic intensity that the data center will face and let the number of servers be this quantity plus its square root.

Let us adopt, in this section, a scaling of the kind that Halfin and Whitt proposed. More precisely, in order to fix ideas, we will simply let

$$\lambda_m = m\mu - \nu\sqrt{m},$$

where ν is some positive constant. Under this choice of scaling, the transition rates of the many-server queue $\hat{X}_m$, with m servers, are given by

$$\hat{q}^m_{n,n+1} = \lambda_m = m\left(\mu - \frac{\nu}{\sqrt{m}}\right) \quad \text{and} \quad \hat{q}^m_{n,n-1} = \mu\min(m,n) = m\mu\min\left(1, \frac{n}{m}\right).$$

We now choose an open set E containing $[0, +\infty)$ and we set

$$\gamma_1(x) = \mu, \quad \gamma_{-1}(x) = \mu\min(1,x) \quad \delta^m_1(x) = -\frac{\nu}{\sqrt{m}} \quad \text{and} \quad \delta^m_{-1}(x) = 0$$

Letting $x = n/m$ we obtain the intensities of $X_m = \hat{X}_m/m$, namely

$$\begin{aligned} q^m_{x,x+m^{-1}} &= m\gamma^m_1(x) + m\delta^m_1(x) = m\beta^m_1(x) \quad \text{and} \\ q^m_{x,x-m^{-1}} &= m\gamma_{-1}(x) + m\delta^m_{-1}(x) = m\beta^m_{-1}(x). \end{aligned}$$

Note that X_m is confined to $[0, +\infty)$ if the initial condition lies there. By Definition 2.1.4 we see that the drift and perturbing drift of this density dependent family are

$$F(x) = \mu\max(0, 1-x) \quad \text{and} \quad G_m(x) = -\frac{\nu}{\sqrt{m}} \quad \forall\ m \geq 1.$$

Let x be the solution, starting at some $x_0 \geq 0$, to the ODE (2.6); in this case

$$\dot{x} = \mu\max(0, 1-x). \tag{3.1}$$

By Theorem 2.2.5 this is the fluid limit of $\{X_m\}_{m\geq 1}$. More precisely, if $X_m(0) \to x_0$ as $m \to \infty$, then we have

$$\sup_{t\in[0,T]} ||X_m(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as } m \to \infty \quad \forall\ T \geq 0.$$

Recall that X_m is the ratio between jobs and servers in the many-server queue with m servers, hence the meaning of the limit x is the ratio between jobs and servers in the fluid scale. The equilibria of equation (3.1) are all the points in $[1, +\infty)$, which represent systems where there are more jobs than servers. Moreover, $x_0 < 1$ implies $x(t) \to 1$ as $t \to +\infty$. This is all reasonable because we are letting $\lambda_m/m\mu \to 1$, and thus the queues $\hat{X}_m$ work closer to the border of their capacity as $m \to \infty$.

In order to estimate the probability that all the servers in the data center are busy, we need to take a closer look; to this end we could try to use Theorem 2.3.2 to compute a diffusion that approximately describes the system's behavior. However, there are two reasons why this is not possible, specifically, the following hypothesis of Theorem 2.3.2 do not hold.

1. F is not differentiable at $y = 1$.
2. $\sqrt{m}\delta_1^m(y) = -\nu$ for all $y \in [0, +\infty)$, thus $\sqrt{m}\delta_1^m$ does not vanish as $m \to \infty$.

In the following sections we will extend Theorem 2.3.2 to contemplate these two situations. Namely, we will prove central limit theorems for families with non-differentiable drifts, and for families whose perturbing drifts do not vanish in the diffusion scale. Before we do that, let us explain what the diffusion approximation should look like in this particular case.

Consider the equilibrium point $x^* = 1$ of equation (3.1), which corresponds to a system where the number of jobs and servers is the same. Also, consider the processes $Z_m = \sqrt{m}(X_m - x^*)$, assume that $Z_m(0) \to Z(0)$ as $m \to \infty$, for some $Z(0) \in \mathbb{R}$, and as we did in Section 2.3 write

$$\begin{aligned} Z_m(t) = Z_m(0) + \sqrt{m}\Sigma_m(t) + \int_0^t \sqrt{m} G_m(X_m(\tau))d\tau \\ + \int_0^t \sqrt{m}\left[F(X_m(\tau)) - F(x^*)\right] d\tau \quad \forall\ t \geq 0; \end{aligned} \tag{3.2}$$

in fact here $F(x^*) = 0$. Recall that $\sqrt{m}\Sigma_m$ is given in terms of two independent and centered Poisson processes of unitary intensity, Y_1 and Y_{-1}, by the expression

$$\sqrt{m}\Sigma_m(t) = \frac{1}{\sqrt{m}} Y_1\left(\int_0^t m\beta_1^m(X_m(\tau))d\tau\right) - \frac{1}{\sqrt{m}} Y_{-1}\left(\int_0^t m\beta_{-1}^m(X_m(\tau))d\tau\right).$$

The first hurdle that we encounter, when we try to reproduce the arguments of Section 2.3, is that we are no longer able to use the series expansion of F to make Z_m appear in the right-hand side of equation (3.2). Indeed, the procedure of Section 2.3 fails because F is not differentiable at $x^* = 1$. Nevertheless, the lateral derivatives of F exist at this point, and this allows to define the map $\partial F : \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$\partial F(y) = \frac{\partial F^-(x^*)}{\partial x} y\mathbb{1}_{y<0} + \frac{\partial F^+(x^*)}{\partial x} y\mathbb{1}_{y\geq 0} = -\mu y\mathbb{1}_{y<0},$$

where the two terms in the middle expression denote, respectively, the left and right lateral derivatives of F. The key features of this map are:

1. It is positively homogeneous in the sense that $\partial F(\alpha y) = \alpha \partial F(y)$ for all $\alpha \geq 0$ and for all $y \in \mathbb{R}$.
2. It is Lipschitz; in fact it is piecewise linear in this particular example.
3. The remainder $R(y) = F(y) - F(x^*) - \partial F(y - x^*)$ satisfies the condition

$$\lim_{y \to x^*} \frac{R(y)}{y - x^*} = 0;$$

as a matter of fact $R \equiv 0$ in this case.

The first of these properties allows to write

$$\begin{aligned}\sqrt{m}\left[F(X_m(\tau)) - F(x^*)\right] &= \sqrt{m}\partial F\left(X_m(\tau) - x^*\right) + \sqrt{m}R\left(X_m(\tau)\right) \\ &= \partial F(Z_m(\tau)) + \sqrt{m}R\left(X_m(\tau)\right).\end{aligned}$$

The above in turn yields an equation that is similar to equation (2.9), namely

$$Z_m(t) = Z_m(0) + U_m(t) + \delta_m(t) + \int_0^t \partial F\left(Z_m(\tau)\right) d\tau \quad \forall\ t \geq 0, \tag{3.3}$$

where, as in Section 2.3, U_m and δ_m are the processes

$$U_m(t) = \sqrt{m}\Sigma_m(t) \quad \text{and} \quad \delta_m(t) = \int_0^t \sqrt{m}\left[R(X_m(\tau)) + G_m(X_m(\tau))\right] d\tau.$$

A key difference is that the integrand $\partial F(Z_m)$ is not linear on Z_m. In Section 2.3, this allowed to explicitly construct a continuous mapping $\phi : D_{\mathbb{R}}[0,T] \longrightarrow D_{\mathbb{R}}[0,T]$ for each $T \geq 0$, such that $\phi_f - f$ was continuous and

$$\phi_f(t) = f(t) + \int_0^t \partial F(\phi_f(\tau))d\tau \quad \forall\ t \in [0,T],$$

for each $f \in D_{\mathbb{R}}[0,T]$. In particular, we had $Z_m = \phi(Z_m(0) + U_m + \delta_m)$. An explicit construction is not possible in this case. However, as we will see in Section 3.3, it is possible to prove that a map ϕ with the latter characteristics exists.

The subsequent step to prove Theorem 2.3.2, after we had constructed ϕ, was to show that $U_m + \delta_m \Rightarrow U$ in $D_{\mathbb{R}}[0,T]$ as $m \to \infty$, where

$$U(t) = W_1\left(\int_0^t \gamma_1(x^*)d\tau\right) - W_{-1}\left(\int_0^t \gamma_{-1}(x^*)d\tau\right) = W_1(\gamma_1(x^*)t) - W_{-1}(\gamma_{-1}(x^*)t)$$

for some independent standard Wiener processes W_1 and W_{-1}. Recall that the strategy was to show that $U_m \Rightarrow U$ in $D_{\mathbb{R}^d}[0,T]$ and

$$\sup_{t\in[0,T]} |\delta_m(t)| \xrightarrow{\mathbb{P}} 0 \quad \text{as } m \to \infty.$$

The latter is no longer true in this case, because the maps $\sqrt{m}G_m$ converge uniformly to the non-zero constant $-\nu$. However, the integral of $\sqrt{m}R(X_m(\tau))$ in the definition of δ_m vanishes, in fact $R \equiv 0$ as we commented before. This results in

$$\sup_{t\in[0,T]} |\delta_m(t) + \nu t| \xrightarrow{\mathbb{P}} 0 \quad \text{as } m \to \infty.$$

Consequently, we actually have the limit $U_m(t) + \delta_m(t) \Rightarrow U(t) - \nu t$, and after using the continuous mapping theorem this results in $Z_m \Rightarrow Z$ in $D_{\mathbb{R}}[0,T]$, where

$$\begin{aligned}Z(t) &= Z(0) + U(t) - \nu t + \int_0^t \partial F(Z(\tau))d\tau \\ &= Z(0) + U(t) + \int_0^t \partial F(Z(\tau)) - \nu d\tau \quad \forall\ t \in [0,T];\end{aligned}$$

note the differences in the integrand, in comparison with equation (2.10). Moreover, since $Z_m \Rightarrow Z$ holds in $D_{\mathbb{R}}[0,T]$ for all $T \geq 0$, then this also holds in $D_{\mathbb{R}}[0,\infty)$.

As in Theorem 2.3.2, the limit process Z may also be regarded as the solution to an SDE. Specifically, Z solves the equation

$$dZ_t = A(Z_t)dt + BdW_t,$$

where W is a unidimensional Wiener process, and A and B are, respectively, the following scalar function and constant.

$$A(y) = \begin{cases} -\mu y - \nu & \text{if } y < 0, \\ -\nu & \text{if } y > 0, \end{cases} \quad \text{and} \quad B = \sqrt{2\mu}.$$

The proof in [14], of this limit in distribution, uses some criteria due to Stone for the convergence of suitably normalized birth-death processes to a diffusion process; these criteria are summarized in [15, Theorem 3.2]. Furthermore, the steady-state Z_∞ of the solutions to the last SDE is computed in [14], and its density is half normal and half exponential:

$$p(y) = (1-\alpha)\frac{\varphi(\beta + y)}{\Phi(\beta)}\mathbb{1}_{y<0} + \alpha\beta e^{-\beta y}\mathbb{1}_{y>0}; \tag{3.4}$$

here φ is the density of the standard normal distribution, Φ is its cumulative distribution function, $\beta = \nu/\mu$ and $\alpha \in (0,1)$ is the probability that $Z_\infty > 0$, which may be expressed in terms of β. The meaning of $Z_\infty > 0$ is that there are more jobs than servers in the system, whereas $Z_\infty < 0$ means that there are more servers than jobs, and these two scenarios have positive probability. This is the distinctive feature of the Halfin-Whitt scaling.

The above computation, of the steady-sate distribution of Z, is carried out in [14] without using the SDE at all; Halfin and Whitt find the stationary distribution of the many-server queue explicitly, and then obtain the steady-state distribution of Z after taking the limit as $m \to \infty$. In Section 3.5 we will instead find the stationary distribution of Z directly from the SDE.

The organization of the subsequent sections is as follows. First, we prove in Section 3.2 a central limit theorem for density dependent families whose perturbing drifts are not negligible in the diffusion scale, but still have a differentiable drift. Afterwards, we prove a central limit theorem, around an equilibrium solution to the fluid dynamics, for density dependent families with a non-differentiable drift, and also non-negligible perturbing drifts. To this end, we study in Section 3.3 solutions to integral equations with a càdlàg input and a Lipschitz field, that is with the form of equation (3.3). The results in this section will help us prove the existence of the map ϕ that we used in the above example, and this will lead to the announced central limit theorem, for families with a non-differentiable drift; this theorem will be proven in Section 3.4. Finally, in Section 3.5, we will discuss the steady-state of some switched diffusions, which appear when we use the results of Section 3.4.

3.2 Generalization of the central limit theorem

Consider an open set $E \subset \mathbb{R}^d$, a finite set of directions $D \subset \mathbb{Z}^d$ and families of non-negative maps $\{\beta_l^k\}_{l\in D}$ with domain E, of the form

$$\beta_l^k = \gamma_l + \delta_l^k.$$

We will consider the density dependent family of continuous time Markov chains X_k that is given by the above maps according to Definition 2.1.2. Moreover, we will continue under the following hypothesis.

Assumption 3.2.1. The maps γ_l are locally Lipschitz. Also, for each compact set $K \subset E$, the maps δ_l^k satisfy the two following conditions.

$$\sup_{x\in K} |\delta_l^k(x)| < \infty \quad \forall\ l \in D,\ k \geq 1 \quad \text{and}$$
$$\lim_{k\to\infty} \sup_{x\in K} k^\alpha |\delta_l^k(x)| = 0 \quad \forall\ l \in D,\ \alpha \in [0, 1/2).$$

Note that Theorem 2.3.1 is still true under this assumption, although in Section 2.3 we assumed that the last of these conditions also held for $\alpha = 1/2$. This was only needed to prove the central limit theorem of Chapter 2.

The drift and perturbing drifts are defined as in Section 2.1, respectively:

$$F(x) = \sum_{l\in D} l\gamma_l(x) \quad \text{and} \quad G_k(x) = \sum_{l\in D} l\delta_l^k(x).$$

Assumption 3.2.1 implies that the drift is locally Lipschitz, and thus Theorem 2.2.5 holds. Namely, suppose that there exists $x_0 \in E$ such that $X_k(0) \to x_0$ as $k \to \infty$, and let x be the solution to the initial value problem $\dot{x} = F(x)$, starting at x_0 and defined in $[0, T]$. Then we have

$$\sup_{t\in[0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty.$$

We will now consider the processes $Z_k = \sqrt{k}(X_k - x)$, that are defined on $[0, T]$ and describe the fluctuations of the processes X_k around their fluid limit x. As in Section 2.3, we will assume that there exists $Z(0) \in \mathbb{R}^d$ such that $Z_k(0) \to Z(0)$ as $k \to \infty$, and we will show that the processes Z_k converge weakly in $D_{\mathbb{R}^d}[0, T]$. To this end, we will adopt the following assumptions.

Assumption 3.2.2. Suppose that the drift F is continuously differentiable. Moreover, assume that there exists a continuous field $G : E \longrightarrow \mathbb{R}^d$ such that

$$\lim_{k\to\infty} \sup_{x\in K} \left|\left|\sqrt{k}G_k(x) - G(x)\right|\right| = 0$$

holds inside of each compact set $K \subset E$. This hypothesis, regarding the perturbing drifts, will determine the appearance of an extra term in the SDE that we wrote in the statement of the central limit theorem of Chapter 2.

In order to prove that the processes Z_k have a limit in distribution, we first

observe that these processes satisfy the equations

$$Z_k(t) = Z_k(0) + U_k(t) + \delta_k(t) + \int_0^t F'(x(\tau))Z_k(\tau)d\tau \quad \forall\, t \in [0,T], \tag{3.5}$$

which we had already introduced in Section 2.3; see equation (2.9). Recall that the processes U_k and δ_k are given by the expressions

$$U_k(t) = \sum_{l\in D} \frac{l}{\sqrt{k}} Y_l\left(\int_0^t k\beta_l^k(X_k(\tau))d\tau\right) \quad \text{and}$$
$$\delta_k = \int_0^t \sqrt{k}\left[R_\tau(X_k(\tau)) + G_k(X_k(\tau))\right]d\tau,$$

where $\{Y_l\}_{l\in D}$ is and independent family of centered Poisson processes with unitary intensity, and $R_\tau(y) = F(y) - F(x(\tau)) - F'(x(\tau))\,(y - x(\tau))$.

By the results in Section 2.3, we know that there exists a continuous mapping $\phi : D_{\mathbb{R}^d}[0,T] \longrightarrow D_{\mathbb{R}^d}[0,T]$ such that the image of $f \in D_{\mathbb{R}^d}[0,T]$ is the unique function ϕ_f such that $\phi_f - f$ is continuous and

$$\phi_f(t) = f(t) + \int_0^t F'(x(\tau))\phi_f(\tau)d\tau \quad \forall\, t \in [0,T].$$

In particular $Z_k = \phi(Z_k(0) + U_k + \delta_k)$; note that since Z_k is a càdlàg function, then the integral in the right-hand side of equation (3.5) is continuous as a function of its upper limit, and hence $Z_k - Z_k(0) - U_k - \delta_k$ is continuous as well.

To prove that the processes Z_k have a limit in distribution, we will first show that the processes $Z_k(0) + U_k + \delta_k$ converge weakly in $D_{\mathbb{R}^d}[0,T]$, and then use the continuous mapping theorem.

To this purpose, we will consider the process

$$U(t) = \sum_{l\in D} lW_l\left(\int_0^t \gamma_l(x(\tau))d\tau\right),$$

where $\{W_l\}_{l\in D}$ is an independent family of standard Wiener processes, and we will prove that the following limit holds.

Theorem 3.2.3. Under the above assumptions $U_k \Rightarrow U$ in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$.

Proof. Consider the processes

$$\tilde{U}_k(t) = \sum_{l\in D} \frac{l}{\sqrt{k}} Y_l\left(\int_0^t k\gamma_l(x(\tau))d\tau\right).$$

By Theorem B.2.1 we know that $\tilde{U}_k \Rightarrow U$ in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$. As a result, by Theorem A.1.7 it is enough to show that

$$\sup_{t\in[0,T]} \left|\left|U_k(t) - \tilde{U}_k(t)\right|\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as } k \to \infty,$$

and to do this it suffices to prove that

$$\sup_{t\in[0,T]} \left|\frac{Y_l(kI_l^k(t)) - Y_l(kJ_l(t))}{\sqrt{k}}\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as } k \to \infty \quad \forall\, l \in D,$$

where I_l^k and J_l are defined, respectively, as follows.

$$I_l^k(t) = \int_0^t \beta_l^k(X_k(\tau))d\tau \quad \text{and} \quad J_l(t) = \int_0^t \gamma_l(x(\tau))d\tau.$$

Fix some $l \in D$ and $\varepsilon > 0$. Since γ_l is locally Lipschitz and $\{x(t) : t \in [0,T]\}$ is compact, there exist $M, \rho > 0$ such that M is a Lipschitz constant for γ_l in the set $\Gamma_\rho = \{y \in E : ||y - x(t)|| \leq \rho \text{ for some } t \in [0,T]\}$; we may choose ρ so that this set is compact. Now choose some $\alpha \in (0, 1/2)$ and consider the random variables

$$\Delta_k = \max\left\{ \sup_{t\in[0,T]} k^\alpha MT\,||X_k(t) - x(t)||\,, \sup_{y\in\Gamma_\rho} k^\alpha T|\delta_l^k(y)| \right\}.$$

Assumption 3.2.1 implies that Theorem 2.3.1 holds and $\Delta_k \xrightarrow{\mathbb{P}} 0$ as $k \to \infty$. Thus, if we fix $0 < \Delta \leq 2MT\rho$, then the probability of the sets $\Omega_k = \{\omega \in \Omega : 2\Delta_k(\omega) > \Delta\}$ converges to zero as $k \to \infty$.

Note that $\omega \in \Omega_k^c$ implies that $||X_k(\omega,t) - x(t)|| \leq (MT)^{-1}\Delta_k(\omega) \leq \rho$ for all $t \in [0,T]$, and hence we have the following inequality for all $\omega \in \Omega_k^c$.

$$\begin{aligned}
|kI_l^k(\omega,t) - kJ_l(t)| &\leq \int_0^t k|\gamma_l(X_k(\omega,\tau)) - \gamma_l(x(\tau))| + k|\delta_l^k(X_k(\omega,\tau))|d\tau \\
&\leq \int_0^t kM\,||X_k(\omega,\tau) - x(\tau)|| + k|\delta_l^k(X_k(\omega,\tau))|d\tau \\
&\leq 2k^{1-\alpha}\Delta_k(\omega) \leq k^{1-\alpha}\Delta.
\end{aligned}$$

It is now convenient to introduce the following notation.

$$\begin{aligned}
A_k &= \left\{\omega \in \Omega : \sup_{t\in[0,T]} \left|\frac{Y_l(kI_l^k(t)) - Y_l(kJ_l(t))}{\sqrt{k}}\right| \geq \varepsilon\right\} \quad \text{and} \\
B_k &= \left\{\omega \in \Omega : \sup_{t\in[0,T],\delta\in[0,\Delta]} \left|\frac{Y_l(kJ_l(t) + k^{1-\alpha}\delta) - Y_l(kJ_l(t))}{\sqrt{k}}\right| \geq \varepsilon\right\}.
\end{aligned}$$

We would like to show that $\mathbb{P}(A_k) \to 0$ as $k \to \infty$ and, to this end, it is enough to prove that $\mathbb{P}(B_k) \to 0$ as $k \to \infty$. Indeed, since $A_k \cap \Omega_k^c \subset B_k \cap \Omega_k^c$, then

$$\mathbb{P}(A_k) \leq \mathbb{P}(\Omega_k) + \mathbb{P}(B_k \cap \Omega_k^c) \leq \mathbb{P}(\Omega_k) + \mathbb{P}(B_k).$$

Let us introduce the notation:

$$\tilde{\Delta}_k = k^{-\alpha}\Delta \quad \text{and} \quad \tilde{Y}_l(t) = Y_l(kt).$$

Using the above notation we may write

$$\sup_{t\in[0,T],\delta\in[0,\Delta]} \left|\frac{Y_l(kJ_l(t) + k^{1-\alpha}\delta) - Y_l(kJ_l(t))}{\sqrt{k}}\right| = \sup_{s\in[0,S],\tilde{\delta}\in[0,\tilde{\Delta}_k]} \left|\frac{\tilde{Y}_l\left(s + \tilde{\delta}\right) - \tilde{Y}_l(s)}{\sqrt{k}}\right|,$$

where $S = J_l(T)$; note that J_l is continuous and non-decreasing with $J_l(0) = 0$, and thus the image of $[0,T]$ under J_l is $[0,S]$.

We will derive a bound for the numerator on the right. In order to do this, let

$S_k = S + \tilde{\Delta}_k$ and consider partitions $\{0 = s_0^k < \cdots < s_{r_k}^k = S_k\}$ such that

$$\frac{\tilde{\Delta}_k}{2} \leq s_{n+1}^k - s_n^k \leq \tilde{\Delta}_k \quad \forall\, n = 0, \ldots, r_k - 1.$$

Given $s \in [0, S]$ and $\theta \in [0, \tilde{\Delta}_k]$ we have three possible scenarios.

1. $s_n^k \leq s \leq s + \theta \leq s_{n+1}^k$ for some $n = 0, \ldots, r_k - 1$.
2. $s_n^k \leq s < s_{n+1}^k < s + \theta \leq s_{n+2}^k$ for some $n = 0, \ldots, r_k - 2$.
3. $s_n^k \leq s < s_{n+1}^k < s_{n+2}^k < s + \theta \leq s_{n+3}^k$ for some $n = 0, \ldots, r_k - 3$.

We will derive the following inequality assuming that s and θ are as in the third case, but similar computations are possible in the two remaining cases.

$$\begin{aligned}\left|\tilde{Y}_l(s+\theta) - \tilde{Y}_l(s)\right| &\leq \left|\tilde{Y}_l(s+\theta) - \tilde{Y}_l(s_{n+2}^k)\right| + \left|\tilde{Y}_l(s_{n+2}^k) - \tilde{Y}_l(s_{n+1}^k)\right| \\ &\quad + \left|\tilde{Y}_l(s_{n+1}^k) - \tilde{Y}_l(s_n^k)\right| + \left|\tilde{Y}_l(s_n^k) - \tilde{Y}_l(s)\right| \\ &\leq 4 \max_{0 \leq n < r_k} \sup_{\tilde{\delta} \in [0, \tilde{\Delta}_k]} \left|\tilde{Y}_l\left(s_n^k + \tilde{\delta}\right) - \tilde{Y}_l(s_n^k)\right|.\end{aligned}$$

Using the bound that we have just computed we obtain the following inequality.

$$\begin{aligned}\mathbb{P}(B_k) &= \mathbb{P}\left(\sup_{s \in [0,S], \tilde{\delta} \in [0, \tilde{\Delta}_k]} \left|\frac{\tilde{Y}_l\left(s + \tilde{\delta}\right) - \tilde{Y}_l(s)}{\sqrt{k}}\right| \geq \varepsilon\right) \\ &\leq \sum_{n=0}^{r_k - 1} \mathbb{P}\left(\sup_{\tilde{\delta} \in [0, \tilde{\Delta}_k]} \left|\frac{\tilde{Y}_l\left(s_n^k + \tilde{\delta}\right) - \tilde{Y}_l(s_n^k)}{\sqrt{k}}\right| \geq \frac{\varepsilon}{4}\right) \\ &= r_k \mathbb{P}\left(\sup_{\delta \in [0, \Delta]} \left|\frac{Y_l(k^{1-\alpha}\delta)}{\sqrt{k}}\right| \geq \frac{\varepsilon}{4}\right).\end{aligned}$$

Since $s_n^k - s_{n-1}^k \geq \tilde{\Delta}_k / 2$ for all $k = 1, \ldots, r_k$, then we have

$$r_k \leq \frac{2S_k}{\tilde{\Delta}_k} = \frac{2k^\alpha S}{\Delta} + 2 \leq 2k^\alpha \left(\frac{S}{\Delta} + 1\right).$$

Thus, applying Doob's maximal inequality to the submartingale Y_l^4, we see that

$$\begin{aligned}r_k \mathbb{P}\left(\sup_{\delta \in [0, \Delta]} \left|\frac{Y_l(k^{1-\alpha}\delta)}{\sqrt{k}}\right| \geq \frac{\varepsilon}{4}\right) &\leq r_k \left(\frac{4}{\varepsilon}\right)^4 \frac{k^{1-\alpha}\Delta + 3\left(k^{1-\alpha}\Delta\right)^2}{k^2} \\ &\leq 2\left(\frac{S}{\Delta} + 1\right)\left(\frac{4}{\varepsilon}\right)^4 \left(\frac{\Delta}{k} + \frac{3\Delta^2}{k^\alpha}\right),\end{aligned}$$

The right-hand side of this equation converges to zero as $k \to \infty$, and thus $\mathbb{P}(B_k) \to 0$ as $k \to \infty$. This completes the proof. □

By Proposition A.1.8, to establish that

$$Z_k(0) + U_k(t) + \delta_k(t) \Rightarrow Z(0) + U(t) + \int_0^t G(x(\tau))d\tau \quad \text{in } D_{\mathbb{R}^d}[0, T] \quad \text{as } k \to \infty,$$

it only remains to prove the following lemma.

Lemma 3.2.4. Under the above assumptions we have

$$\sup_{t\in[0,T]}\left|\left|\delta_k(t)-\int_0^t G(x(\tau))d\tau\right|\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as } k\to\infty.$$

Proof. Using the definition of δ_k, given below equation (3.5), we may write

$$\begin{aligned}\sup_{t\in[0,T]}\left|\left|\delta_k(t)-\int_0^t G(x(\tau))d\tau\right|\right| \leq T \sup_{t\in[0,T]}\sqrt{k}\,||R_t(X_k(t))|| \\ + T\sup_{t\in[0,T]}\left|\left|\sqrt{k}G_k(X_k(t))-G(x(t))\right|\right|\end{aligned} \tag{3.6}$$

To begin, we will deal with the second term on the right-hand side:

$$\begin{aligned}\sup_{t\in[0,T]}\left|\left|\sqrt{k}G_k(X_k(t))-G(x(t))\right|\right| \leq \sup_{t\in[0,T]}\left|\left|\sqrt{k}G_k(X_k(t))-G(X_k(t))\right|\right| \\ + \sup_{t\in[0,T]}||G(X_k(t))-G(x(t))||\,.\end{aligned}$$

Consider the set $\Gamma_\rho = \{y\in E : ||y-x(t)||\leq\rho \text{ for some } t\in[0,T]\}$, which is compact for any sufficiently small constant $\rho>0$. By Theorem 2.2.5, we know that for each ω, outside some fixed null set, there exists $k_0(\omega)$ such that $k\geq k_0(\omega)$ implies $X_k(\omega,t)\in\Gamma_\rho$ for all $t\in[0,T]$. Hence, Assumption 3.2.2 implies that the first term on the right-hand side converges to zero almost surely as $k\to\infty$. Moreover, since G is uniformly continuous in Γ_ρ, then the second term also converges to zero almost surely as $k\to\infty$ by Theorem 2.2.5.

Now it only remains to be shown that the first term on the right-hand side of equation (3.6) converges to zero in probability as $k\to\infty$. Before we do that, since the remainder $R_t(y)\to 0$ faster than $||y-x(t)||$ as $y\to x(t)$, we may agree on defining the next expression as zero at $y=x(t)$, namely

$$\frac{R_t(y)}{||y-x(t)||}=0 \quad \text{at } y=x(t).$$

Under this convention it is possible to write the following inequality.

$$\begin{aligned}\sup_{t\in[0,T]}\sqrt{k}\,||R_t(X_k(t))|| &= \sup_{t\in[0,T]}\sqrt{k}\,||X_k(t)-x(t)||\,\frac{||R_t(X_k(t))||}{||X_k(t)-x(t)||} \\ &\leq \sup_{t\in[0,T]}\sqrt{k}\,||X_k(t)-x(t)||\sup_{t\in[0,T]}\frac{||R_t(X_k(t))||}{||X_k(t)-x(t)||}.\end{aligned}$$

We will now make use of Lemma 2.2.4 and the definitions therein. Using the notation in this lemma, we may write for each ω outside of the null set N, the following inequality.

$$\sup_{t\in[0,T]}\sqrt{k}\,||X_k(\omega,t)-x(t)||\leq\sqrt{k}\varepsilon_k(\omega)e^{MT} \quad \forall\ k\geq k_0(\omega).$$

Let A_k be the set of those $\omega \in \Omega$ for which

$$\sup_{t\in[0,T]} \sqrt{n}\,||X_n(\omega,t) - x(t)|| \leq \sqrt{n}\varepsilon_n(\omega)e^{MT} \tag{3.7}$$

for all $n \geq k$. These sets increase to a set that contains N^c, and thus has probability one. Also, given $\varepsilon > 0$, consider the set

$$B_k = \left\{\omega \in \Omega : \sqrt{k}\varepsilon_k(\omega) \sup_{t\in[0,T]} \frac{||R_t(X_k(\omega,t))||}{||X_k(\omega,t) - x(t)||} \geq \frac{\varepsilon}{e^{MT}}\right\}.$$

Since equation (3.7) holds inside A_k, then we have

$$\mathbb{P}\left(T \sup_{t\in[0,T]} \sqrt{k}\,||R_t(X_k(t))|| \geq \varepsilon\right) \leq \mathbb{P}(A_k^c) + \mathbb{P}(A_k \cap B_k) \leq \mathbb{P}(A_k^c) + \mathbb{P}(B_k).$$

We already know that $\mathbb{P}(A_k^c) \to 0$ as $k \to \infty$. Thus, it will be enough to show that $\mathbb{P}(B_k) \to 0$ as $k \to \infty$. In other words, we would like to prove that

$$\sqrt{k}\varepsilon_k \sup_{t\in[0,T]} \frac{||R_t(X_k(t))||}{||X_k(t) - x(t)||} \xrightarrow{\mathbb{P}} 0 \quad \text{as } k \to \infty.$$

Since F is continuously differentiable, then we know that F is uniformly differentiable on any compact set K. By this we mean that for all $\varepsilon > 0$, there exists $\delta > 0$ with the following property: if $||z - y|| < \delta$ and the line segment $[y, z]$ is contained inside of K, then

$$\frac{||R_y(z)||}{||z - y||} = \frac{||F(z) - F(y) - F'(y)(z - y)||}{||y - z||} < \varepsilon;$$

we refer the reader to Proposition E.2.1. Note that given any sufficiently small constant $\rho > 0$, the set $\Gamma_\rho = \{y \in E : ||y - x(t)|| \leq \rho \text{ for some } t \in [0, T]\}$ is compact and has the property that $||y - x(t)|| < \rho$ implies $[x(t), y] \subset \Gamma_\rho$. As a result, the uniform differentiability of F and Theorem 2.2.5 imply that

$$\sup_{t\in[0,T]} \frac{||R_t(X_k(t))||}{||X_k(t) - x(t)||} \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty.$$

Furthermore, recalling from Lemma 2.2.4 the definition of ε_k, we may write

$$\sqrt{k}\varepsilon_k = ||Z_k(0)|| + T \sup_{x\in A} \sqrt{k}\,||G_k(x)|| + \sum_{l\in D} ||l|| \sup_{t\in[0,T]} \frac{\left|Y_l(k\overline{\beta}_l t)\right|}{\sqrt{k}}.$$

On the one hand, we have, by Assumption 3.2.2, the bound

$$\limsup_{k\to\infty} \left[||Z_k(0)|| + T \sup_{x\in A} \sqrt{k}\,||G_k(x)||\right] \leq ||Z(0)|| + T \sup_{x\in A} ||G(x)||,$$

and consequently, we see that

$$\left[||Z_k(0)|| + T \sup_{x\in A} \sqrt{k}\,||G_k(x)||\right] \sup_{t\in[0,T]} \frac{||R_t(X_k(t))||}{||X_k(t) - x(t)||} \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty.$$

On the other hand, consider an independent family $\{W_l\}_{l\in D}$ of standard Wiener

processes. Note that the supremum norm is continuous in the Skorohod topology. Therefore, using Theorem B.2.1, the independence of the family $\{Y_l\}_{l\in D}$ and the continuous mapping theorem, we conclude that

$$\sum_{l\in D} ||l|| \sup_{t\in[0,T]} \frac{|Y_l(k\overline{\beta}_l t)|}{\sqrt{k}} \Rightarrow \sum_{l\in D} ||l|| \sup_{t\in[0,T]} |W_l(\overline{\beta}_l t)| \quad \text{in } \mathbb{R} \quad \text{as } k\to\infty,$$

and using Proposition A.1.9 we may write

$$\left[\sum_{l\in D} ||l|| \sup_{t\in[0,T]} \frac{|Y_l(k\overline{\beta}_l t)|}{\sqrt{k}}\right] \sup_{t\in[0,T]} \frac{||R_t(X_k(t))||}{||X_k(t)-x(t)||} \xrightarrow{\mathbb{P}} 0 \quad \text{as } k\to\infty.$$

This completes the proof. □

Now we are ready to prove the main result of this section.

Theorem 3.2.5. Assume that $Z_k(0) \to Z(0)$ as $k\to\infty$, for some $Z(0)\in\mathbb{R}^d$. Also, suppose that assumptions 3.2.1 and 3.2.2 hold. Then $Z_k \Rightarrow Z$ in $D_{\mathbb{R}^d}[0,T]$ as $k\to\infty$, where Z is the continuous process that satisfies the equation

$$Z(t) = Z(0) + U(t) + \int_0^t F'(x(\tau))Z(\tau) + G(x(\tau))d\tau \quad \forall\, t\in[0,T].$$

Furthermore, Z has the same finite-dimensional distributions as the solution to

$$dZ_t = [A_t Z_t + G(x(t))]\,dt + B_t dW_t, \tag{3.8}$$

where W_t is a d-dimensional Wiener process with independent coordinates, A_t is the Jacobian matrix of F at the point $x(t)$ and B_t is the $d\times d$ matrix

$$B(t) = \sqrt{\sum_{l\in D} ll^T \gamma_l(x(t))}.$$

Here the square root is that of a positive semi-definite matrix. Note that the drift of this SDE has an extra term in comparison with equation (2.13)

Proof. As we have already observed, $Z_k = \phi(Z_k(0)+U_k+\delta_k)$ for all $k\geq 1$. Define

$$\tilde{U}(t) = U(t) + \int_0^t G(x(\tau))d\tau.$$

Theorem 3.2.3 and Lemma 3.2.4 imply that $Z_k(0)+U_k+\delta_k \Rightarrow Z(0)+\tilde{U}$ in $D_{\mathbb{R}^d}[0,T]$ as $k\to\infty$, by Proposition A.1.8.

If $Z=\phi(Z(0)+\tilde{U})$, then the continuity of ϕ implies that $Z_k\Rightarrow Z$ in $D_{\mathbb{R}^d}[0,T]$ as $k\to\infty$, by the continuous mapping theorem. Moreover, the definition of ϕ implies that $Z-\tilde{U}$ is continuous, and thus Z is continuous as well. Also, we have

$$Z(t) = Z(0) + U(t) + \int_0^t F'(x(\tau))Z(\tau) + G(x(\tau))d\tau \quad \forall\, t\in[0,T],$$

again by the definition of ϕ. The link between this equation and the SDE (3.8) is as in Subsection 1.2.2. □

As in Section 2.3 we may also write an explicit equation for Z, namely

$$Z(t) = Z(0) + \tilde{U}(t) + \int_0^t \Phi(\tau, t) F'(x(\tau)) \left[Z(0) + \tilde{U}(\tau) \right] d\tau,$$

where $\tilde{U}$ is as in the proof of the preceding theorem and Φ is the fundamental matrix that solves the initial value problem

$$\frac{\partial \Phi(s,t)}{\partial t} = F'(x(t))\Phi(s,t) \quad \text{and} \quad \Phi(s,s) = \mathrm{Id} \quad \forall\ s,t \in [0,T].$$

This allows to show, by the same arguments of Section 2.3, that Z is a time inhomogeneous Gaussian process.

In this case we see from equation (3.8) that the mean of Z solves the ODE

$$\mu(t) = Z(0) + \int_0^t A(\tau)\mu(\tau) + G(x(\tau)) d\tau,$$

which is different from the corresponding equation of Subsection 2.3.1. However, the centered process $Z - \mu$ solves the SDE

$$d(Z-\mu)_t = A_t (Z-\mu)_t dt + B_t dW_t,$$

which is equation (2.13) from Subsection 2.3.1, and therefore the covariance of Z is exactly as in Proposition 2.3.5, that is

$$\Sigma(s,t) = \int_0^s \Phi(\tau, s) B_\tau B_\tau^T \Phi(\tau, t)^T d\tau \quad \forall\ 0 \leq s \leq t \leq T.$$

3.3 Integral equations with a càdlàg input

A key step for proving the central limit theorems of Chapter 2 and the previous section is the construction of the mapping ϕ, and to that end the hypothesis that F is continuously differentiable is crucial. As a matter of fact, in the case of families with a non-differentiable drift we cannot construct this mapping explicitly.

However, we still may prove that a map with the same properties exists. To this purpose, consider a càdlàg function $f : [0,T] \longrightarrow \mathbb{R}^d$ and a globally Lipschitz field $H : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. In this section we will study equations of the form

$$\varphi(t) = f(t) + \int_0^t H(\varphi(\tau)) d\tau \quad \forall\ t \in [0,T].$$

Definition 3.3.1. We will say that $\varphi : [0,T] \longrightarrow \mathbb{R}^d$ is a solution to the integral equation with input f and field H if φ satisfies the above and $\varphi - f$ is continuous.

Moreover, given some interval $I \subset [0,T]$ and an initial condition $(t_0, x_0) \in I \times \mathbb{R}^d$, we will say that $\varphi : I \longrightarrow \mathbb{R}^d$ is a local solution if $\varphi - f$ is continuous and

$$\varphi(t) = x_0 + f(t) - f(t_0) + \int_{t_0}^t H(\varphi(\tau)) d\tau \quad \forall\ t \in I.$$

The next lemma is the analog of Picard's theorem within the context of the equations that we are considering in this section.

Lemma 3.3.2. There exists $\varepsilon > 0$ such that for all $t_0 \in [0,T]$ and $x_0 \in \mathbb{R}^d$ there exists a unique local solution starting at (t_0, x_0) and defined on $(t_0-\varepsilon, t_0+\varepsilon)\cap[0,T]$.

Proof. Let M be a Lipschitz constant for H and choose $\varepsilon > 0$ such that $M\varepsilon < 1$. Fix $(t_0, x_0) \in [0,T] \times \mathbb{R}^d$, let $I = (t_0 - \varepsilon, t_0 + \varepsilon) \cap [0,T]$ and note that if a local solution exists, then by definition, it must lie in

$$\mathcal{F} = \left\{\varphi : I \longrightarrow \mathbb{R}^d : \varphi - f \text{ is continuous}\right\}.$$

Endow this space with the metric ρ inherited from the supremum norm

$$\rho(\varphi, \psi) = \sup_{t\in I} ||\varphi(t) - \psi(t)|| \quad \forall\ \varphi, \psi \in \mathcal{F}.$$

It is clear that $(\mathcal{F}, \rho)$ is a complete metric space. Furthermore, $\mathcal{F}$ is isometrically isomorphic to the space of continuous functions with domain I. We are going to consider the map $T : \mathcal{F} \longrightarrow \mathcal{F}$ such that

$$T\varphi(t) = x_0 + f(t) - f(t_0) + \int_{t_0}^{t} H(\varphi(\tau))d\tau \quad \forall\ t \in I.$$

Note that φ is càdlàg because f has this property and $\varphi - f$ is continuous. Hence, since H is continuous, $H(\varphi)$ is càdlàg and thus integrable. Moreover, the integral on the right-hand side is continuous as a function of the upper limit, because the integrand is bounded, and therefore $T\varphi - f$ is continuous, in other words $T\varphi \in \mathcal{F}$. The following inequality shows that T is a contraction.

$$\begin{aligned}\rho(T\varphi, T\psi) = \sup_{t\in I} ||T\varphi(t) - T\psi(t)|| &\leq \sup_{t\in I} \left| \int_{t_0}^{t} ||H(\varphi(\tau)) - H(\psi(\tau))||\, d\tau \right| \\ &\leq M\varepsilon \sup_{t\in I} ||\varphi(t) - \psi(t)|| < \rho(\varphi, \psi).\end{aligned}$$

Therefore, the fixed point theorem ensures that there exists a unique $\varphi \in \mathcal{F}$ such that $T\varphi = \varphi$, and this completes the proof of the claim. □

The radius ε of the time interval where local solutions exist and are unique is independent of the initial condition (t_0, x_0). This happens because H is uniformly Lipschitz, and it will help us prove the following theorem.

Theorem 3.3.3. There exists a unique solution $\varphi : [0,T] \longrightarrow \mathbb{R}^d$ to

$$\varphi(t) = f(t) + \int_{0}^{t} H(\varphi(\tau))d\tau \quad \forall\ t \in [0,T]. \tag{3.9}$$

Proof. Let $\varepsilon > 0$ be as in the statement of Lemma 3.3.2 and choose a partition $0 = t_0 < \cdots < t_n = T$ such that $t_{i+1} - t_i < \varepsilon$ for all $i = 0, \ldots, n-1$. Furthermore, let $I_i = (t_i - \varepsilon, t_i + \varepsilon) \cap [0,T]$ and define inductively $\varphi_i : I_i \longrightarrow \mathbb{R}^d$ to be the unique local solution to the corresponding equation of the ones below.

$$\varphi_0(t) = f(0) + f(t) - f(0) + \int_{0}^{t} H(\varphi_0(\tau))d\tau \quad \forall\ t \in I_0,$$

$$\varphi_i(t) = \varphi_{i-1}(t_i) + f(t) - f(t_i) + \int_{t_i}^{t} H(\varphi_i(\tau))d\tau \quad \forall\ t \in I_i.$$

Consider the map $\varphi : [0,T] \longrightarrow \mathbb{R}^d$ such that $\varphi(t) = \varphi_i(t)$ for all $t \in [t_i, t_{i+1}]$, since $\varphi_{i-1}(t_i) = \varphi_i(t_i)$, this map is well defined. Moreover, it is easy to check that φ solves equation (3.9). The uniqueness of φ follows from the uniqueness of the solutions φ_i in the intervals I_i for all $i = 0, \dots, n-1$. □

For some fixed Lipschitz field $H : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, the previous theorem allows to define a map $\phi : D_{\mathbb{R}^d}[0,T] \longrightarrow D_{\mathbb{R}^d}[0,T]$ such that the image of $f \in D_{\mathbb{R}^d}[0,T]$ is the unique solution ϕ_f to equation (3.9) when f is the input. In other words, ϕ_f is the unique function such that $\phi_f - f$ is continuous and

$$\phi_f(t) = f(t) + \int_0^t H(\phi_f(\tau))d\tau \quad \forall\, t \in [0,T].$$

Theorem 3.3.4. Assume that $H(0) = 0$, then the map $\phi : D_{\mathbb{R}^d}[0,T] \longrightarrow D_{\mathbb{R}^d}[0,T]$ is continuous in the Skorohod topology.

Proof. Let $M > 0$ be a Lipschitz constant for H. Then $||H(x)|| \leq M\,||x||$ because $H(0) = 0$. As a result, we have the following for each $f \in D_{\mathbb{R}^d}[0,T]$.

$$\begin{aligned} ||\phi_f(t)|| &\leq ||f(t)|| + \int_0^t ||H(\phi_f(\tau))||\, d\tau \\ &\leq ||f(t)|| + \int_0^t M\,||\phi_f(\tau)||\, d\tau \quad \forall t \in [0,T], \end{aligned}$$

and now the boundedness of f, and Gronwall's inequality, yield a uniform bound for $||\phi_f(t)||$ that we will denote K_f, specifically

$$\sup_{t\in[0,T]} ||\phi_f(t)|| \leq e^{MT} \sup_{t\in[0,T]} ||f(t)|| = K_f.$$

Consider some $g \in D_{\mathbb{R}^d}[0,T]$ and suppose that $\delta > d_0(f,g)$. Recall that the metric d_0, that we used in Lemma 2.3.4 and is defined in Appendix A, generates the Skorohod topology. Moreover, remember that the condition $d_0(f,g) < \delta$ implies that there exists some increasing continuous bijection $\lambda : [0,T] \longrightarrow [0,T]$ such that

$$\sup_{0\leq s<t\leq T} \left|\log\left(\frac{\lambda(t)-\lambda(s)}{t-s}\right)\right| \leq \delta \quad \text{and} \quad \sup_{t\in[0,T]} ||f(\lambda(t)) - g(t)|| \leq \delta.$$

The first inequality implies that λ is differentiable almost everywhere in $[0,T]$ and its derivative satisfies $|\lambda'(t) - 1| \leq e^{\delta} - 1$ at the points where it is defined, thus

$$\begin{aligned} \phi_f(\lambda(t)) - \phi_g(t) &= f(\lambda(t)) - g(t) + \int_0^{\lambda(t)} H(\phi_f(\tau))d\tau - \int_0^t H(\phi_g(\tau))d\tau \\ &= f(\lambda(t)) - g(t) + \int_0^t H(\phi_f(\lambda(s)))\lambda'(s)ds - \int_0^t H(\phi_g(s))ds \\ &= f(\lambda(t)) - g(t) + \int_0^t H(\phi_f(\lambda(s)))[\lambda'(s) - 1]ds \\ &+ \int_0^t H(\phi_f(\lambda(s))) - H(\phi_g(s))ds \quad \forall\, t \in [0,T], \end{aligned}$$

and we have the following inequality for all $t \in [0,T]$.

$$\begin{aligned}||\phi_f(\lambda(t)) - \phi_g(t)|| &\leq ||f(\lambda(t)) - g(t)|| + \int_0^t M\,||\phi_f(\lambda(s))||\left(e^{\delta}-1\right)ds \\ &\quad + \int_0^t M\,||\phi_f(\lambda(s)) - \phi_g(s)||\,ds \\ &\leq \delta + MK_f\left(e^{\delta}-1\right)T + \int_0^t M\,||\phi_f(\lambda(s)) - \phi_g(s)||\,ds.\end{aligned}$$

Finally, using Gronwall's inequality we obtain the bound

$$\sup_{t\in[0,T]} ||\phi_f(\lambda(t)) - \phi_g(t)|| \leq \left[\delta + MK_f\left(e^{\delta}-1\right)T\right]e^{MT}.$$

This holds for all $\delta > d_0(f,g)$, and therefore we may write

$$d_0(\phi_f,\phi_g) \leq \left[d_0(f,g) + MK_f\left(e^{d_0(f,g)}-1\right)T\right]e^{MT}.$$

For some fixed $f \in D_{\mathbb{R}^d}[0,T]$, the expression on the right converges to zero as $d_0(f,g) \to 0$, and this implies that ϕ is continuous at f. □

The previous theorem is also true without the assumption $H(0) = 0$; the proof is almost the same, although slightly messier. However, we will only apply this theorem to fields such that $H(0) = 0$.

3.4 Refinement for non-differentiable drifts

In this section we consider density dependent families whose drifts are not differentiable, and we prove a central limit theorem in the case where the nominal solution to the fluid dynamics (2.6) is an equilibrium point.

As in the previous section, we will consider an open set $E \subset \mathbb{R}^d$, a finite set of directions $D \subset \mathbb{R}^d$ and a family $\{\beta_l^k\}_{l\in D}$ of non-negative maps with domain E, again of the form $\beta_l^k = \gamma_l + \delta_l^k$. We will further consider the density dependent family of continuous time Markov chains X_k that the above maps define.

We will suppose that Assumption 3.2.1 holds, as in Section 3.2. Recall that this implies that the drift of the family is locally Lipschitz, and therefore we know that the strong law of large numbers of Section 2.2 holds. Furthermore, we will assume that the ODE $\dot{x} = F(x)$ admits some equilibrium point $x^* \in E$, and we will suppose that $X_k(0) \to x^*$ as $k \to \infty$. Then, by Theorem 2.2.5, we know that

$$\sup_{t\in[0,T]} ||X_k(t) - x^*|| \xrightarrow{\text{a.s.}} 0 \quad \text{as } k \to \infty \quad \forall\, T \geq 0.$$

Now we may consider the process $Z_k = \sqrt{k}(X_k - x^*)$ that describes the fluctuations of X_k around the fluid equilibrium x^*. To begin we will fix some $T \geq 0$ and let these processes be defined in $[0,T]$, but we will get rid of this restriction afterwards. As in the previous section, we will assume that there exists some $Z(0) \in \mathbb{R}^d$ such that $Z_k(0) \to Z(0)$ as $k \to \infty$. Our goal is to show that the processes Z_k have a

limit in distribution in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$, and to this end we will continue under the following hypothesis.

Assumption 3.4.1. Suppose that there exists a Lipschitz field $\partial F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ with the following properties.

1. ∂F is positively homogeneous, in the sense that $\partial F(\alpha x) = \alpha \partial F(x)$ for all $x \in \mathbb{R}^d$ and for all $\alpha \geq 0$.

2. The remainder $R(x) = F(x) - F(x^*) - \partial F(x - x^*)$ is such that
$$\lim_{x \to x^*} \frac{||R(x)||}{||x - x^*||} = 0.$$

As the reader may have noticed, the Lipschitz field ∂F plays the role of the drift's differential at the equilibrium point x^*. Moreover, this field is easy to construct when the drift is piecewise differentiable around x^*. In order to provide this construction, let us introduce the notation $I_- = (-\infty, 0)$, $I_+ = [0, +\infty)$ and $J = \{-, +\}^d$. Also, consider a basis $\{v_1, \ldots, v_d\}$ of $\mathbb{R}^d$, such that $||v_i|| = 1$ for all $i = 1, \ldots, d$, and note that we have the decomposition
$$\mathbb{R}^d = \bigcup_{j \in J} \{x^* + I_{j_1} v_1 \times \cdots \times I_{j_d} v_d\}.$$
Consider the lateral directional derivatives at x^*, along $v_1, \ldots, v_d$, specifically
$$\frac{\partial F^-(x^*)}{\partial v_i} = \lim_{h \to 0^+} \frac{F(x^* - h v_i) - F(x^*)}{h}, \quad \frac{\partial F^+(x^*)}{\partial v_i} = \lim_{h \to 0^+} \frac{F(x^* + h v_i) - F(x^*)}{h}.$$
We may now prove the following.

Proposition 3.4.2. Suppose that for each $j \in J$ there exists a differentiable field $F_j : E \longrightarrow \mathbb{R}^d$ such that $F(x) = F_j(x)$ for all $x \in E \cap \overline{\{x^* + I_{j_1} v_1 \times \cdots \times I_{j_d} v_d\}}$. Given $v \in \mathbb{R}^d$ consider the unique decomposition $v = \alpha_1 v_1 + \cdots + \alpha_d v_d$ and define
$$\partial F(v) = \sum_{i=1}^{d} \left[\frac{\partial F^-(x^*)}{\partial v_i} \alpha_i \mathbb{1}_{\alpha_i < 0} + \frac{\partial F^+(x^*)}{\partial v_i} \alpha_i \mathbb{1}_{\alpha_i \geq 0} \right] v_i.$$
Then Assumption 3.4.1 holds.

Proof. It is easy to check that ∂F is positively homogeneous and Lipschitz; each of the terms in the sum that defines ∂F have these properties.

Therefore, we only need to check that the remainder $R(x)$ converges to zero faster than $||x - x^*||$ as $x \to x^*$, namely we must prove that
$$\lim_{x \to x^*} \frac{||F(x) - F(x^*) - \partial F(x - x^*)||}{||x - x^*||} = 0.$$
Note that ∂F agrees with the differential of F_j at x^* on $I_{j_1} v_1 \times \cdots \times I_{j_d} v_d$. Also, for each $x \in E$ there exists $i \in J$ such that $x - x^* \in I_{i_1} v_1 \times \cdots \times I_{i_d} v_d$, and hence we

may write the inequality

$$\frac{||F(x) - F(x^*) - \partial F(x - x^*)||}{||x - x^*||} = \frac{||F(x) - F(x^*) - F_i'(x^*)(x - x^*)||}{||x - x^*||}$$
$$\leq \max_{j \in J} \frac{\left|\left|F(x) - F(x^*) - F_j'(x^*)(x - x^*)\right|\right|}{||x - x^*||}.$$

The right-hand side converges to zero as $x \to x^*$ and this completes the proof. $\square$

Returning to the proof of the weak convergence of the processes Z_k, we will need, as in Section 3.2, the following hypothesis.

Assumption 3.4.3. There exists a continuous field $G : E \longrightarrow \mathbb{R}^d$ such that

$$\lim_{k \to \infty} \sup_{x \in K} \left|\left| \sqrt{k} G_k(x) - G(x) \right|\right| = 0$$

holds inside of each compact set $K \subset E$.

By the positive homogeneity of ∂F we have

$$\sqrt{k}\,[F(X_k(t)) - F(x^*)] = \partial F(Z_k(t)) + \sqrt{k} R(X_k(t)),$$

which allows to write an equation for Z_k that is very similar to the one that we used in sections 2.3 and 3.2, namely

$$Z_k(t) = Z_k(0) + U_k(t) + \delta_k(t) + \int_0^t \partial F(Z_k(\tau)) d\tau \quad \forall\ t \in [0, T]. \tag{3.10}$$

If we compare with equations (2.9) and (3.5), in this case $\partial F(Z_k(\tau))$ replaces $F'(x^*) Z_k(\tau)$ in the integral that appears on the right-hand side. Also, recall that

$$U_k(t) = \sum_{l \in D} \frac{l}{\sqrt{k}} Y_l \left(\int_0^t k \beta_l^k(X_k(\tau)) d\tau \right) \quad \text{and}$$
$$\delta_k = \int_0^t \sqrt{k}\,[R(X_k(\tau)) + G_k(X_k(\tau))]\, d\tau,$$

where $\{Y_l\}_{l \in D}$ is and independent family of centered Poisson processes with unitary intensities; these are the same expressions that appeared in sections 2.3 and 3.2.

The fact that ∂F is Lipschitz is what allows to use the results of Section 3.3, which tell us that there exists a continuous $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ such that

$$\phi_f(t) = f(t) + \int_0^t \partial F(\phi_f(\tau)) d\tau \quad \forall\ t \in [0, T]$$

and $\phi_f - f$ is continuous, for all $f \in D_{\mathbb{R}^d}[0, T]$. Probably the main difference with the central limit theorems of sections 2.3 and 3.2 is that, in the context of this section, it is no longer possible to construct ϕ explicitly. Because of this, the results of Section 3.3 are crucial.

Since Z_k is a càdlàg function and ∂F is continuous, then the integral on the right-hand side of equation (3.10) is continuous as a function of its upper limit, and hence $Z_k - Z_k(0) - U_k - \delta_k$ is continuous as well. Consequently, $Z_k = \phi(Z_k(0) + U_k + \delta_k)$ and as in Section 3.2 we now want to show that $U_k + \delta_k$ has a limit in distribution,

so that we can afterwards use the continuous mapping theorem to prove that the processes Z_k themselves have a limit in distribution as well.

As in Section 3.2, the processes U_k converge weakly to

$$U(t) = \sum_{l \in D} lW_l \left(\int_0^t \gamma_l(x^*) d\tau \right) = \sum_{l \in D} lW_l \left(\gamma_l(x^*) t \right),$$

where $\{W_l\}_{l \in D}$ is an independent family of standard Wiener processes. Indeed, the reader may check that the proof of Theorem 3.2.3 only relies on Assumption 3.2.1 and the hypothesis $Z_k(0) \to Z(0)$ as $k \to \infty$.

Also, the claim of Lemma 3.2.4 is still true in the context of this section, namely

$$\sup_{t \in [0,T]} \left\| \delta_k(t) - \int_0^t G(x^*) d\tau \right\| \xrightarrow{\mathbb{P}} 0 \quad \text{as } k \to \infty.$$

The proof follows from the same arguments that we used in Section 3.2, using the hypothesis that we stated in Assumption 3.4.1 for the remainder R.

We are now ready to prove the central limit theorem that we were seeking. The proof will be as in Section 3.2 with the difference that we will now be able to prove weak convergence in $D_{\mathbb{R}^d}[0, \infty)$, rather than only in $D_{\mathbb{R}^d}[0, T]$ for some fixed $T \geq 0$.

Theorem 3.4.4. Assume that $Z_k(0) \to Z(0)$ as $k \to \infty$, for some $Z(0) \in \mathbb{R}^d$. Also, suppose that assumptions 3.2.1, 3.4.1 and 3.4.3 hold. Then $Z_k \Rightarrow Z$ in $D_{\mathbb{R}^d}[0, \infty)$ as $k \to \infty$, where Z is the continuous process that satisfies the equation

$$Z(t) = Z(0) + U(t) + \int_0^t \partial F(Z(\tau)) + G(x^*) d\tau \quad \forall\ t \geq 0.$$

Furthermore, Z has the same finite-dimensional distributions as the solution to

$$dZ_t = \left[\partial F(Z_t) + G(x^*) \right] dt + B dW_t, \tag{3.11}$$

where W is a d-dimensional Wiener process, with independent coordinates, and B is the square root of the following positive semi-definite symmetric matrix.

$$B = \sqrt{\sum_{l \in D} l l^T \gamma_l(x^*)}.$$

Proof. For each $n \geq 1$ let $\phi^n : D_{\mathbb{R}^d}[0, n] \longrightarrow D_{\mathbb{R}^d}[0, n]$ be the continuous function such that the image of $f \in D_{R^d}[0, n]$ is the unique $\phi^n_f \in D_{\mathbb{R}^d}[0, n]$ such that $\phi^n_f - f$ is continuous and

$$\phi^n_f(t) = f(t) + \int_0^t \partial F(\phi^n_f(\tau)) d\tau \quad \forall\ t \in [0, n].$$

Consider now the process

$$\tilde{U}(t) = U(t) + tG(x^*).$$

For each $n \geq 1$ define $Z^n = \phi^n(Z(0) + \tilde{U})$; here we are in fact considering the restriction of $\tilde{U}$ to $[0, n]$. The processes Z^n are continuous and $m \leq n$ implies $Z^m(t) = Z^n(t)$ for all $t \in [0, m]$. Therefore, the process Z such that $Z(t) = Z^n(t)$

for all $t \in [0, n]$ is well defined, continuous and satisfies the equation

$$Z(t) = Z(0) + U(t) + \int_0^t \partial F(Z(\tau)) + G(x^*)d\tau \quad \forall\, t \geq 0.$$

If we now fix some $T \geq 0$, then $Z_k(0) + U_k + \delta_k \Rightarrow Z(0) + \tilde{U}$ in $D_{\mathbb{R}^d}[0, T]$ as $k \to \infty$. Hence, by the continuous mapping theorem $Z_k \Rightarrow Z$ in $D_{\mathbb{R}^d}[0, T]$ as well. Since this is true for all $T \geq 0$, then the convergence in distribution also holds in $D_{\mathbb{R}^d}[0, \infty)$ by Theorem A.3.6. □

Even though the SDE (3.11) is similar to that of Section 3.2, the fact that the drift is no longer affine on Z complicates the characterization of solutions. For instance, solutions to this SDE will in general not be time inhomogeneous Gaussian processes, as we will see in the next section.

In the sequel we discuss the problem of finding the steady-state of solutions to SDEs with the form of equation (3.11); we more precisely assume that ∂F is piecewise linear, which is the case when ∂F is constructed using Proposition 3.4.2. Finding the steady-state of solutions to equation (3.11) is of particular interest if we want to use this equation to characterize the typical behavior of a system within any application.

3.5 The steady-state of some switched diffusions

We begin by providing some background and specifying the type of processes that we will consider. To this purpose, let $a : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ and $b : \mathbb{R}^d \longrightarrow S_d^+$ be coefficients satisfying the hypothesis of Theorem D.2.2, for the existence and uniqueness of solutions to SDEs; here S_d^+ is the space of symmetric positive semi-definite matrices. Consider now the Feller diffusion X associated to the SDE

$$dX_t = a(X_t)dt + b(X_t)dW_t. \tag{3.12}$$

If X admits an exponentially ergodic invariant measure, as in Definition D.3.3, then the steady-state of solutions to equation (3.12) exists and it is distributed according to this measure. The Foster-Lyapunov criteria of Section D.3 is useful for proving exponential ergodicity, but it does not characterize the invariant measure. We would like to know, for instance, if this measure is absolutely continuous with respect to the Lebesgue measure; and in that case we would like to compute its density.

Suppose that $\{T_t\}_{t \geq 0}$ is the Feller semigroup of operators defined by the transition function P of the Feller process X, specifically

$$T_t f(x) = \int_{\mathbb{R}^d} f(y) P_t(x, dy) = \mathbb{E}_x[f(X_t)]$$

for all $x \in \mathbb{R}^d$ and $f \in C_0(\mathbb{R}^d)$.

An initial distribution π of X is said to be an invariant measure if

$$\int_{\mathbb{R}^d} P_t(x, \Gamma)\pi(dx) = \pi(\Gamma)$$

for all $t \geq 0$ and all Borel sets $\Gamma \subset \mathbb{R}^d$. In terms of the semigroup of operators $\{T_t\}_{t\geq 0}$, the latter condition is equivalent to

$$\int_{\mathbb{R}^d} T_t f(x) - f(x)\pi(dx) = 0 \quad \forall\ f \in C_0(\mathbb{R}^d).$$

Furthermore, if we let A be the infinitesimal generator of $\{T_t\}_{t\geq 0}$, and we pick some function f in the domain $\mathcal{D}(A)$ of A, then by definition we have

$$\lim_{t\to 0} \sup_{x\in\mathbb{R}^d} \left|\left| \frac{T_t f(x) - f(x)}{t} - Af(x) \right|\right| = 0.$$

Therefore, using this uniform convergence, we see that the invariance of π implies

$$\int_{\mathbb{R}^d} Af(x)\pi(dx) = \lim_{t\to 0} \frac{1}{t} \int_{\mathbb{R}^d} T_t f(x) - f(x)\pi(dx) = 0.$$

The converse is also true by the first item of Proposition C.1.6, which implies that for each $f \in C_0(\mathbb{R}^d)$ there exists some $g \in \mathcal{D}(A)$ such that $T_t f - f = Ag$. Hence, a probability measure π is invariant for X if and only if

$$\int_{\mathbb{R}^d} Af(x)\pi(dx) = 0 \quad \forall\ f \in \mathcal{D}(A). \tag{3.13}$$

By the observations at the end of Appendix C, we know that $C_c^\infty(\mathbb{R}^d) \subset \mathcal{D}(A)$. Furthermore, if we let $\sigma^2 = bb^T$, then A agrees in $C_c^\infty(\mathbb{R}^d)$ with the second order differential operator $L : C_c^\infty(\mathbb{R}^d) \longrightarrow C_c^\infty(\mathbb{R}^d)$ such that

$$L\varphi = \sum_{i=1}^d a_i \frac{\partial \varphi}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^d \sigma_{i,j}^2 \frac{\partial^2 \varphi}{\partial x_i \partial x_j} \quad \forall\ \varphi \in C_c^\infty(\mathbb{R}^d).$$

In fact $C^2(\mathbb{R}^d) \subset \mathcal{D}(A)$ and $A\varphi$ is given by the above expression for all $\varphi \in C^2(\mathbb{R}^d)$. As a result of the previous observation, equation (3.13) implies that

$$\int_{\mathbb{R}^d} L\varphi(x)\pi(dx) = 0 \quad \forall\ \varphi \in C_c^\infty(\mathbb{R}^d), \tag{3.14}$$

which is called a weak elliptic equation for measures. Note that this is a weaker statement than that of equation (3.13), and consequently it does not necessarily imply that π is an invariant measure for X. The point is that the space $C_c^\infty(\mathbb{R}^d)$ may be much smaller than $\mathcal{D}(A)$. Nevertheless, equation (3.14) may help us find potential invariant measures. With this in mind we introduce the following definition.

Definition 3.5.1. A multi-index is $\alpha \in \mathbb{N}^d$ and we let $|\alpha| = \alpha_1 + \cdots + \alpha_d$. In addition, given any $\varphi \in C_c^\infty(\mathbb{R}^d)$ we define

$$\partial^\alpha \varphi = \frac{\partial^{|\alpha|}\varphi}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \quad \forall\ \alpha \in \mathbb{N}^d.$$

We say that a locally integrable function f has weak α-derivative of order $|\alpha|$ if there exists a locally integrable function $\partial^\alpha f$ such that

$$\int_{\mathbb{R}^d} f(x)\partial^\alpha \varphi(x)dx = (-1)^{|\alpha|} \int_{\mathbb{R}^d} \partial^\alpha f(x)\varphi(x)dx \quad \forall\ \varphi \in C_c^\infty(\mathbb{R}^d).$$

Proposition E.3.1 implies that weak α-derivatives are unique.

Let μ be a Borel probability measure satisfying equation (3.14) and assume that μ has a density p with respect to the Lebesgue measure; in fact under fairly weak hypothesis it is proven in [3, Chapter 1.1] that this happens. Furthermore, suppose that $a_i p$ and $\sigma_{i,j}^2 p$ have weak derivatives up to the second order for all $i, j \in \{1, \dots, d\}$. Then we may write equation (3.14) in terms of the weak derivatives of $a_i p$ and $\sigma_{i,j}^2 p$. Specifically, for each $\varphi \in C_c^\infty(\mathbb{R}^d)$ we have

$$\begin{aligned}\int_{\mathbb{R}^d} L\varphi(x)\mu(dx) &= \int_{\mathbb{R}^d} \left[\sum_{i=1}^d a_i(x)\frac{\partial\varphi(x)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^d \sigma_{i,j}^2(x)\frac{\partial^2\varphi(x)}{\partial x_i \partial x_j}\right] p(x)dx \\ &= \int_{\mathbb{R}^d} \left[-\sum_{i=1}^d \partial^i \left[a_i(x)p(x)\right] + \frac{1}{2}\sum_{i,j=1}^d \partial^{i,j}\left[\sigma_{i,j}^2(x)p(x)\right]\right]\varphi(x)dx\end{aligned}$$

Equation (3.14) tells us that the left-hand side is equal to zero for all $\varphi \in C_c^\infty(\mathbb{R}^d)$, then by Proposition E.3.1 we know that

$$-\sum_{i=1}^d \partial^i (a_i p) + \frac{1}{2}\sum_{i,j=1}^d \partial^{i,j}\left(\sigma_{i,j}^2 p\right) = 0 \quad \text{a.e.} \tag{3.15}$$

Furthermore, reversing the above procedure it is easy to see that μ solves the weak elliptic equation (3.14) whenever p satisfies equation (3.15), which is known as the Fokker-Planck equation or the forward Kolmogorov equation.

It is important to stress that, in general, solutions to equation (3.15) only give potential solutions to equation (3.13). However, there are cases where solutions to equation (3.14) are unique and may be found by solving the Fokker-Planck equation. In these cases, if we know that a solution to equation (3.13) exists, then the measure that arises from the Fokker-Planck equation is the unique invariant measure.

3.5.1 Piecewise affine unidimensional SDEs

Consider a unidimensional SDE with the form of equation (3.11) when ∂F is piecewise linear. Specifically, suppose that $\sigma^2 > 0$ is constant and that

$$a(x) = \begin{cases} \alpha^- x + \beta & \text{if } x < 0, \\ \alpha^+ x + \beta & \text{if } x \geq 0. \end{cases}$$

In the unidimensional case is possible to prove that equation (3.14) has at most one solution, even when a is just a locally integrable function; see [3, Proposition 1.6.2]. We will use the Fokker-Planck equation to find the solution when a is as above.

Assume that p is twice differentiable, except at $x = 0$, with locally integrable derivatives; we may rewrite equation (3.15) as follows.

$$\begin{aligned}\frac{\sigma^2}{2}\frac{\partial^2 p(x)}{\partial x^2} - \frac{\partial[(\alpha^- x + \beta)p(x)]}{\partial x} = 0 \quad \forall\, x < 0 \quad \text{and} \\ \frac{\sigma^2}{2}\frac{\partial^2 p(x)}{\partial x^2} - \frac{\partial[(\alpha^+ x + \beta)p(x)]}{\partial x} = 0 \quad \forall\, x > 0.\end{aligned} \tag{3.16}$$

Each of these equations is a second order homogeneous ODE, in particular the

space of solutions of either of the above equations has dimension two. Let us focus on solving the second equation; the analysis is analogous for the other one.

Case I

First, suppose that $\alpha^+ \neq 0$. Let $\mu = -\beta/\alpha^+$ and $\nu = -\sigma^2/2\alpha^+$, then we have

$$\nu \frac{\partial^2 p(x)}{\partial x^2} + \frac{\partial[(x-\mu)p(x)]}{\partial x} = 0 \quad \forall\, x > 0.$$

Note that p solves this equation if and only if there exists $c_1 \in \mathbb{R}$ such that

$$\nu \frac{\partial p(x)}{\partial x} + (x-\mu)p(x) = c_1 \quad \forall\, x > 0.$$

The general solution to this first order ODE is

$$p(x) = e^{-\frac{(x-\mu)^2}{2\nu}} \left[\frac{c_1}{\nu} \int_0^x e^{\frac{(t-\mu)^2}{2\nu}} dt + c_2 \right].$$

Since p is the density of a probability measure, we need it to be non-negative and to integrate one over the real line. In particular the integral of p over $(0, +\infty)$ must be finite. If $\nu < 0$ this requires that c_1 and c_2 are such that

$$\lim_{x \to +\infty} \frac{c_1}{\nu} \int_0^x e^{\frac{(t-\mu)^2}{2\nu}} dt + c_2 = 0,$$

otherwise p would not vanish at infinity. However, p is not integrable even in the latter case: using L'Hôpital's rule we see that $p(x)$ decays as $1/x$ as $x \to +\infty$.

$$\lim_{x \to +\infty} xp(x) = \lim_{x \to +\infty} xe^{-\frac{(x-\mu)^2}{2\nu}} \left[\frac{c_1}{\nu} \int_0^x e^{\frac{(t-\mu)^2}{2\nu}} dt + c_2 \right] = c_1 > 0.$$

The last inequality follows from the fact that $c_1 < 0 < c_2$. It is clear that c_1 and c_2 must have different sign for p to vanish at infinity. Also, if c_1 was positive and c_2 negative, then $p(x)$ would be negative for all sufficiently small $x > 0$.

In the case $\nu > 0$ a similar analysis shows that c_1 has to be zero. Here the integral in the definition of p diverges when $x \to +\infty$, and when $c_1 \neq 0$ this allows to use L'Hôpital's rule to show again that $p(x)$ decays as $1/x$ as $x \to +\infty$. Hence, we see that p is as below, it has the form of a Gaussian kernel.

$$p(x) = c_2 e^{-\frac{(x-\mu)^2}{2\nu}}.$$

Case II

Suppose now that $\alpha^+ = 0$, in this case we let $\nu = -\sigma^2/2\beta$ and we see that

$$\nu \frac{\partial^2 p(x)}{\partial x^2} + \frac{\partial p(x)}{\partial x} = 0 \quad \forall\, x > 0.$$

Any solution p to the last equation solves the first order ODE

$$\nu \frac{\partial p(x)}{\partial x} + p(x) = c_1 \quad \forall\, x > 0,$$

In this case the general solution is

$$p(x) = e^{-\frac{x}{\nu}} \left[\frac{c_1}{\nu} \int_0^x e^{\frac{t}{\nu}} dt + c_2 \right] = c_1 + (c_2 - c_1) e^{-\frac{x}{\nu}}.$$

For $\nu < 0$ the last expression is not integrable over $(0, +\infty)$ unless $p \equiv 0$, whereas in the case $\nu > 0$ we must take $c_1 = 0$ so that p vanishes at infinity. This yields that p is as follows, it has the form of an exponential kernel.

$$p(x) = c_2 e^{-\frac{x}{\nu}}.$$

Summarizing, some twice differentiable probability density function p solves equation (3.16) if and only if the two following conditions hold.

1. $\alpha^- < 0$ or $\alpha^- = 0$ and $\beta > 0$.

2. $\alpha^+ < 0$ or $\alpha^+ = 0$ and $\beta < 0$.

Furthermore, in that case $p(x) = p^-(x)\mathbb{1}_{x<0} + p^+(x)\mathbb{1}_{x>0}$, where

$$p^-(x) = \begin{cases} c^- e^{\frac{-(x-\mu^-)^2}{2\nu_1^-}} & \text{if } \alpha^- < 0, \\ c^- e^{-\frac{x}{\nu_2}} & \text{if } \alpha^- = 0,\ \beta > 0; \end{cases}$$

$$p^+(x) = \begin{cases} c^+ e^{\frac{-(x-\mu^+)^2}{2\nu_1^+}} & \text{if } \alpha^+ < 0, \\ c^+ e^{-\frac{x}{\nu_2}} & \text{if } \alpha^+ = 0,\ \beta < 0; \end{cases}$$

here the c^- and c^+ are such that p integrates one over the real line, also

$$\mu^i = -\frac{\beta}{\alpha^i}, \quad \nu_1^i = -\frac{\sigma^2}{2\alpha^i} \quad \text{and} \quad \nu_2 = -\frac{\sigma^2}{2\beta} \quad \forall\ i \in \{-, +\}.$$

In other words p results from pasting two densities, which may be Gaussian or exponential depending on the coefficients of the SDE.

Note that p may solve equation (3.16) and still not solve equation (3.15). Nonetheless, if the weak derivatives of p were

$$\partial^1 p(x) = \frac{\partial p^-(x)}{\partial x}\mathbb{1}_{x<0} + \frac{\partial p^+(x)}{\partial x}\mathbb{1}_{x>0} \quad \text{and}$$
$$\partial^2 p(x) = \frac{\partial^2 p^-(x)}{\partial x^2}\mathbb{1}_{x<0} + \frac{\partial^2 p^+(x)}{\partial x^2}\mathbb{1}_{x>0},$$

then it is clear that p would solve equation (3.15). In order to determine when this happens, pick some $\varphi \in C_c^\infty(\mathbb{R})$ and compute

$$\begin{aligned} \int_{-\infty}^{+\infty} p(x)\frac{\partial \varphi(x)}{\partial x} dx &= \int_{-\infty}^{0} p^-(x)\frac{\partial \varphi(x)}{\partial x} dx + \int_0^{+\infty} p^+(x)\frac{\partial \varphi(x)}{\partial x} dx \\ &= -\int_{-\infty}^{0} \frac{\partial p^-(x)}{\partial x}\varphi(x) dx - \int_0^{+\infty} \frac{\partial p^+(x)}{\partial x}\varphi(x) dx \\ &\quad + [p^-(0) - p^+(0)]\varphi(0). \end{aligned}$$

This means that we must have $p^-(0) = p^+(0)$ if we want $\partial^1 p$ to be as above, in other

words p has to be continuous; note that this gives another equation for c^- and c^+. Moreover, if we use the integration by parts formula again, we get

$$\begin{aligned}\int_{-\infty}^{+\infty} p(x)\frac{\partial^2\varphi(x)}{\partial x^2}dx &= \int_{-\infty}^{0} p^-(x)\frac{\partial^2\varphi(x)}{\partial x^2}dx + \int_{0}^{+\infty} p^+(x)\frac{\partial^2\varphi(x)}{\partial x^2}dx \\ &= \int_{-\infty}^{0} \frac{\partial^2 p^-(x)}{\partial x^2}\varphi(x)dx + \int_{0}^{+\infty} \frac{\partial^2 p^+(x)}{\partial x^2}\varphi(x)dx \\ &\quad + \left[p^-(0) - p^+(0)\right]\frac{\partial\varphi(0)}{\partial x} + \left[\frac{\partial p^+(0)}{\partial x} - \frac{\partial p^-(0)}{\partial x}\right]\varphi(0).\end{aligned}$$

Thus, we also need p to have a continuous derivative if we want $\partial^2 p$ to be as above. However, the reader may check that this is automatic once that we have imposed the condition $p^-(0) = p^+(0)$; this is a consequence of $\mu^i\nu_2 + \nu_1^i = 0$ for all $i \in \{-, +\}$.

Note that the set of equations

$$\begin{cases} c^- \displaystyle\int_{-\infty}^{0} p^-(x)dx + c^+ \int_{0}^{+\infty} p^+(x)dx = 1, \\ p^+(0) - p^-(0) = 0, \end{cases}$$

completely determines the constants c^- and c^+. The resulting p is a probability density function that is twice differentiable, except at $x = 0$, and is a solution to equation (3.15). Recall that equation (3.14) has at most one solution in the current setting by [3, Proposition 1.6.2], thus $\pi(dx) = p(x)dx$ is this solution.

Moreover, it is easy to see that the SDE with coefficients a and σ^2 admits a unique and exponentially ergodic invariant measure; the reader may check that the Foster-Lyapunov function $V(x) = x^2$ satisfies the hypothesis of Theorem D.3.4. Since π is the unique solution to equation (3.14), then π is this measure.

Remark 3.5.2. The SDE that appears at the end of Section 3.1 is a particular case of the piecewise affine unidimensional SDEs that we are considering here. The density of its invariant measure, which appears in equation (3.4), corresponds to the half-Gaussian half-exponential case.

3.5.2 Comments on the multidimensional case

Finding the invariant measure of X is much more difficult in the multidimensional case. Indeed, the first hurdle that we encounter is that solutions to equation (3.14) need not be unique when $d > 1$. As a matter of fact, equation (3.14) may admit several solutions even in the case where a is smooth and σ^2 is constant; the reader may find an example of this in [3, Example 1.6.3].

Another obstacle that we encounter is that the Fokker-Planck equation (3.15) takes the form of a partial differential equation (PDE) in the multidimensional case. Therefore, it is not always possible to completely characterize its solutions, and it may even be difficult to find particular solutions unless the PDE. However, if the PDE is well-known, it may be possible to prove that the Fokker-Planck equation has a unique solution and it may also be possible to find this solution explicitly.

Chapter 4

Dynamic right sizing of computing capacity

4.1 Motivation

The Internet has expanded unceasingly since its beginnings, and it currently hosts numerous online services. The implementation of these services typically requires considerable infrastructure, because of the sheer volume of information that needs to be handled; not in vain the last two decades have seen data centers multiply around the globe to provide the necessary computing resources. These digital factories are not always exploited directly by their owners; because of the high initial costs of infrastructure, many application providers are not owners of data centers, but prefer to avoid startup costs by sharing the computing resources. In this context, landlords of the Internet, such as Amazon and Google, have founded cloud networks: hosts of large numbers of servers that are rented on the fly to businesses worldwide.

A major concern among application providers is that costumers today are highly delay sensitive: a small wait in accessing a service can unfavorably affect the perceived quality of the application, and lead to a decline in usage; with the obvious adverse impact on revenues. For instance, studies show that delaying results to shopping queries in a second may result in e-commerce sales dropping noticeably. Fortunately, the performance of online applications, particularly in terms of latency, may be enhanced by increasing the computing capacity; but this has the obvious drawback of requiring to rent additional servers, in the case of cloud-based service providers, and the disadvantage of higher maintenance costs for data center owners. Indeed, the price of supplying energy to active servers within a data center, and the associated cooling expenditures, comprise a significant fraction of the budget of these facilities.

Hence, a crucial challenge for large scale cloud-based businesses and data centers is to achieve a highly efficient server utilization, that yields excellent user-perceived performance, while using the smallest possible amount of resources. A major compli-

cation is that many applications must deal with uncertain and time-varying demand patterns, which calls for a dynamic right sizing of the active computing capacity. Specifically, this refers to designing an automatic control rule capable of deciding in real time whether there are dispensable servers or, on the contrary, additional capacity needs to be summoned; this decision could be taken, for instance, by assessing the number of pending requests in the system. In cloud-based applications the corresponding action is executed by adjusting the number of instances in use, whereas the implementation in data centers requires servers to transition between active and power-saving modes.

The notion of service elasticity is essential to our hopes of deploying systems with auto-scaling capacity, and lies at the heart of the cloud computing paradigm. This notion hinges on the premise that the vast amount of resources is not likely to act as a bottleneck in any practical sense. Nevertheless, ideal service elasticity does not exist, in the sense that ramping up servers involves a significant time lag, which cannot be ignored. As a result, we must resort to a slight over-provisioning of computing systems if we want to minimize the delay experienced by application users, while we simultaneously cope with the setup lag of servers. An important question, that this work intends to answer, concerns elucidating the extent of the over-provisioning that we need.

In this chapter we perform a mathematical study of the above problem, using tools from queueing theory. Recall that within this framework, the simplest model of a computing system comprises a pool of servers and a single dispatcher; jobs that require to be processed are received sequentially by the dispatcher, and are then sent to an available server. Since servers cannot be spawned instantly, it is necessary to include at least one queue in this model, to hold requests when there are no idle servers to process them.

A very active and recent literature assumes that jobs must be immediately dispatched to a server which, if currently busy, may retain the job in a dedicated queue; for examples on this treatment of the problem see [13,26,27]. However, the approach of this work is to allow the dispatcher to store pending requests in a centralized queue until some server becomes idle; this has been studied for instance in [11, 12, 28, 35]. In this setting it is essential to reduce the number of queued queries to a minimum, not only to ensure that application costumers experience a small latency, but also to maintain the dispatcher's buffer as empty as possible; because storing a large number of queries in a centralized queue may be technologically infeasible. The challenge is to achieve this in a regime where the computing capacity is being dynamically right sized to match the workload. The more general situation where, as in the latter case, a system works close to the limit of its capacity, has been termed heavy traffic in the queueing literature. Understanding the behavior of systems that operate in this setting, but with a fixed number of servers, is a classical problem; for example see [14] and references therein.

4.2 A more realistic infinite-server queue

In the preceding section we posed the problem of right sizing the active capacity of cloud environments and data centers to an unknown external demand; being the best case scenario that of active capacity matching the workload exactly: this prevents the waste of capital in idle capacity and avoids costumers the annoying waits caused by queueing delays. In an idealized setting, where servers can be spawned or deleted instantly, this is achieved by the infinite-server queue, which serves jobs upon their arrival without having to maintain any otiose capacity; in practice there exist, however, non-negligible lags in the creation and deletion of servers. In this chapter we discuss the right sizing of capacity in this context, following the lines of our recent paper [12]. We begin by looking at an analog of the infinite-server queue in a setting where creation and deletion lags are contemplated.

Let us recall the model introduced in Section 1.3 with the latter situation in mind; the provisioning rule that we proposed there was as follows.

- A request is issued to the cloud or data center infrastructure, asking to shut down a server, immediately after each job departure; these requests are executed with an exponential delay of mean $1/c$ seconds.
- If a job arrives in the presence of idle servers, then the job is assigned to one of the idle servers and one of the shut down requests is withdrawn, if there are any of them pending.
- When a job arrives, and has to be queued, a new server is requested, but the infrastructure makes the server available only after an exponential time of mean $1/b$ seconds.
- If in the meanwhile one of the servers becomes idle, then the request is canceled and this idle server takes care of the queued job.

In the above model jobs are supposed to arrive according to a Poisson process of intensity λ jobs per second, and service times are assumed to be exponential with mean $1/\mu$ seconds. The state of the queueing system that we have just described is characterized by the number of servers and jobs, respectively, M and N; we will often use the coordinate notation $X = (M, N)$. As we commented in Section 1.3, the number of pending shut down requests coincides with the number of idle servers $[M - N]^+$, whereas the number of pending server requests is equal to the number of queued jobs $[N - M]^+$. Therefore, the dynamics of this system are given by the transitions diagram of Figure 4.1.

We would like to understand the large scale behavior of the number of jobs and servers, that is with the arrival rate λ approaching infinity; note that λ is a measure of demand, whereas the service rate μ represents the individual capacity of servers, and thus shall remain fixed. The number of jobs is clearly greater than it would be in an ideal infinite-server queue facing the same arrivals, where each job is assigned a server right away upon its arrival. In the latter case, the average occupancy in the

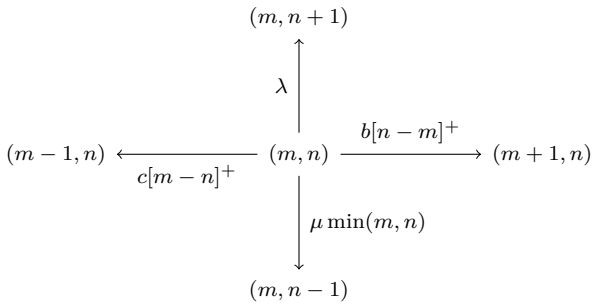


Figure 4.1: Markovian model of a more realistic infinite-server queue, considering the existence of lags in the creation and deletion of servers.

steady-state is equal to the traffic intensity $\rho = \lambda/\mu$, and the fact that this number diverges as $\lambda \to \infty$ justifies the terminology large scale.

Consider then a sequence of systems $\hat{X}_k = (\hat{M}_k, \hat{N}_k)$ facing arrivals at rate $k\lambda$; in any other respect these are identical to the system that we have described above, in particular they behave according to Figure 4.1 with λ replaced by $k\lambda$. The average occupancy in $\hat{X}_k$ is greater than $k\rho$, therefore the sequence has a degenerate limit as $k \to \infty$. This motivates the normalization $X_k = \hat{X}_k/k$; in coordinates we write $X_k = (M_k, N_k)$. The latter yields a density dependent family, generated by the maps

$$\beta_l(m,n) = \begin{cases} b[n-m]^+ & \text{if } l = (1,0), \\ c[m-n]^+ & \text{if } l = -(1,0), \\ \lambda & \text{if } l = (0,1), \\ \mu \min(m,n) & \text{if } l = -(0,1). \end{cases}$$

It is worth pointing out that the perturbations δ_l^k are identically zero and because of that we omit the superscript k when we write β_l.

Before we can apply the results of the last two chapters, we must compute the drift of this density dependent family, which is

$$F(m,n) = \begin{bmatrix} b[n-m]^+ - c[m-n]^+ \\ \lambda - \mu \min(m,n) \end{bmatrix}.$$

This is a Lipschitz field that is not differentiable along the diagonal of the first quadrant. Hence, we may use the strong law of large numbers of Chapter 2, but we must resort to Chapter 3 for a central limit theorem.

First, we compute the fluid limit of the family, which is given by Theorem 2.2.5. At a macroscopic level, this theorem tells us that the behavior of the chains X_k is governed by the ODE

$$\begin{aligned} \dot{m} &= b[n-m]^+ - c[m-n]^+, \\ \dot{n} &= \lambda - \mu \min(m,n). \end{aligned} \tag{4.1}$$

More precisely, suppose that the chains X_k are realized over the same probability space and have deterministic initial conditions that converge to some x_0 lying in the

first quadrant. In that case

$$\lim_{k\to\infty} \sup_{t\in[0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \forall\, T \geq 0,$$

where x is the solution to equation (4.1) that starts at x_0. Figure 4.2 illustrates how the processes X_k are approximated by x when k is large.

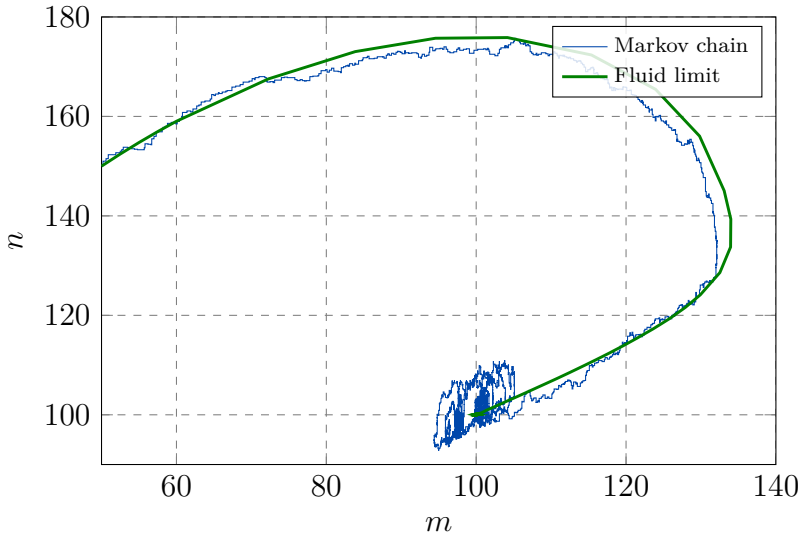


Figure 4.2: Sample path of the Markov chain $X_k(t)$ and the fluid limit $x(t)$; the parameters of the simulation are $\lambda = 100$, $\mu = 1$, $b = 1/2$, $c = 3$, $x_0 = (50, 150)$ and $k = 10$.

As shown in Figure 4.2, regardless of the initial condition, the processes X_k eventually end up hovering around (ρ, ρ), where capacity matches demand.

Proposition 4.2.1. The dynamics (4.1) have a unique equilibrium point x^*, with coordinates $m^* = \rho$ and $n^* = \rho$, and this equilibrium is a global attractor.

Proof. The set $\{(m, n) \in [0, +\infty)^2 : m > n \geq \rho\}$ is invariant under (4.1) because the field points downwards at $\{(m, n) \in [0, +\infty)^2 : m = n > \rho\}$ and the half line $\{(m, n) \in [0, +\infty)^2 : m > \rho, n = \rho\}$ is traveled by solutions. Since the Jacobian matrix of the field in $\{(m, n) \in [0, +\infty)^2 : m > n \geq \rho\}$ has negative eigenvalues, solutions starting in this set remain there forever and approach x^* as $t \to +\infty$.

Consider now the restriction of (4.1) to the set $\{(m, n) \in [0, +\infty)^2 : m \leq n\}$. The linear extension of these dynamics to the entire quadrant would have x^* as a global attractor. Moreover, the field of the dynamics (4.1) points upwards at the line segment $\{(m, n) \in [0, +\infty)^2 : m = n < \rho\}$. Thus, solutions starting inside of $\{(m, n) \in [0, +\infty)^2 : m \leq n\}$ remain in this set forever and approach x^* as $t \to +\infty$, or alternatively fall into $\{(m, n) \in [0, +\infty)^2 : m > n \geq \rho\}$, where they remain and approach x^* as $t \to +\infty$.

Finally, consider the restriction of (4.1) to $\{(m, n) \in [0, +\infty)^2 : m > n, n < \rho\}$; its linear extension to the entire quadrant would have x^* as a global attractor. Hence, solutions starting in $\{(m, n) \in [0, +\infty)^2 : m > n, n < \rho\}$ remain in this set forever

and approach x^* as $t \to +\infty$, or alternatively fall into either of the other already analyzed sets; more precisely, they may only fall into $\{(m,n) \in [0,+\infty)^2 : m \leq n\}$ but this is not relevant for proving the claim. $\square$

The interpretation of the last result, taking into account the strong law of large numbers of Theorem 2.2.5, is that the processes X_k are attracted towards the equilibrium x^* as $t \to +\infty$, especially when k is large; this is depicted in Figure 4.2.

Moreover, by Theorem 2.3.1 we know that

$$\sup_{t\in[0,T]} \left|\left|\hat{X}_k(t) - kx(t)\right|\right| = \sup_{t\in[0,T]} k\,||X_k(t) - x(t)|| = o(k^{1-\alpha}) \quad \text{a.s.} \quad \forall\, \alpha \in [0,1/4).$$

This means that the difference between $\hat{X}_k$ and kx is negligible as $k \to \infty$; it scales sublinearly, while the arrival rate of jobs is scaling linearly.

In order to prove a central limit theorem for the Markov chains X_k, we must resort to Theorem 3.4.4, which requires to construct a field $\partial F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ that satisfies Assumption 3.4.1. Since F is piecewise affine, we may use Proposition 3.4.2 to the latter end. With this in mind, let $\nu = [-1\ 1]^T$, and consider the matrices

$$A_1 = \begin{bmatrix} -b & b \\ -\mu & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} -c & c \\ 0 & -\mu \end{bmatrix},$$

the Jacobians of F in $\{(m,n) \in [0,\infty)^2 : m < n\}$ and $\{(m,n) \in [0,\infty)^2 : m > n\}$, respectively. According to Proposition 3.4.2 we may take

$$\partial F(y) = A_1 y \mathbb{1}_{\langle y,\nu\rangle \geq 0} + A_2 y \mathbb{1}_{\langle y,\nu\rangle < 0}.$$

Now let W be a bidimensional Wiener process and consider the matrix

$$B = \begin{bmatrix} \sqrt{\beta_{(1,0)}(x^*) + \beta_{(-1,0)}(x^*)} & 0 \\ 0 & \sqrt{\beta_{(0,1)}(x^*) + \beta_{(0,-1)}(x^*)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2\lambda} \end{bmatrix}.$$

By Theorem 3.4.4, under suitable hypothesis on the initial conditions, the processes $Z_k = \sqrt{k}(X_k - x^*)$ converge weakly in $D_{\mathbb{R}^2}[0,\infty)$, as $k \to \infty$, to a diffusion Z that solves the SDE

$$dZ_t = \partial F(Z_t)dt + BdW_t.$$

Unfortunately, the non-linear switching in ∂F precludes us from computing the stationary distribution of Z. Still, we may say something about the steady-state of our original system X by just looking at the chain of Figure 4.1. Specifically, we will prove that this chain is positive recurrent, and we will then use this fact to compute the ratio between queue length and over-provisioning in the steady-state.

In order to prove that X is positive recurrent, we will use a classic Foster-Lyapunov criteria for continuous time Markov chains; we state it below.

Theorem 4.2.2. Consider a continuous time Markov chain with state-space S and infinitesimal generator Q. Suppose in addition that there exists $V : S \longrightarrow [0,+\infty)$ with the following properties.

1. There exist positive constants d and e, and a finite set $C \subset S$, such that $QV(y) \leq -d + e\mathbb{1}_C(y)$ for all $y \in S$.

2. The chain is non-explosive or $\{y \in S : V(y) \leq M\}$ is finite for all $M \geq 0$.

Then the chain is positive recurrent.

The state-space of our Markov chain X is the lattice $\mathbb{N}^2$. Thus, $QV(y) \to -\infty$ as $y \to \infty$ implies the first of the above conditions. Similarly, $V(y) \to +\infty$ as $y \to \infty$ implies the second of these conditions. We will show that X is positive recurrent by exhibiting a non-negative function V with these two properties, and to this end we will use the following result. We omit the proof because it is very similar to that of Proposition 4.3.1, which is given in the next section.

Proposition 4.2.3. Suppose that $4b \geq c$. There exists a positive definite symmetric matrix P such that $A_i^T P + P A_i$ is negative definite for all $i \in \{1, 2\}$.

The existence of P implies that $V : \mathbb{R}^2 \longrightarrow \mathbb{R}$ such that $V(y) = (y - x^*)^T P (y - x^*)$ is a common quadratic Lyapunov function for the dynamics (4.1). This means that $V(x^*) = 0$ and that for all $y \neq x^*$ we have $V(y) > 0$ and $\nabla V(y) \partial F(y) < 0$. In particular, this implies Proposition 4.2.1 in the case $4b \geq c$.

To show that V is a Foster-Lyapunov function for X, let Q be the infinitesimal generator of this chain, and let $D = \{\pm(1, 0), \pm(0, 1)\}$, then we have

$$\begin{aligned} QV(y) &= \sum_{l \in D} [V(y + l) - V(y)] \beta_l(y) \\ &= \sum_{l \in D} [(y - x^* + l)^T P (y - x^* + l) - (y - x^*)^T P (y - x^*)] \beta_l(y) \\ &= \sum_{l \in D} [2(y - x^*)^T P l + l^T P l] \beta_l(y) \\ &= 2(y - x^*)^T P F(y) + \sum_{l \in D} l^T P l \beta_l(y) \quad \forall\, y \in \mathbb{R}^2. \end{aligned}$$

The last term on the right-hand side is a piecewise affine function of the coefficients of y, whereas the first term is piecewise quadratic and given by

$$2(y - x^*)^T P F(y) = \begin{cases} (y - x^*)^T (A_1^T P + P A_1)(y - x^*) & \text{if } \langle y, \nu \rangle \geq 0, \\ (y - x^*)^T (A_2^T P + P A_2)(y - x^*) & \text{if } \langle y, \nu \rangle < 0; \end{cases}$$

this results from the identity $F(y) = \partial F(y - x^*)$. Since the above quadratic forms are negative definite, we see that $QV(y) \to -\infty$ as $y \to \infty$; note in addition that $V(y) \to +\infty$ as $y \to \infty$. This proves that X is positive recurrent.

Suppose now that $X_\infty = (M_\infty, N_\infty)$ is distributed according to the invariant measure π of X, and let e be the identity function, then

$$\mathbb{E}[F(X_\infty)] = \mathbb{E}\left[\sum_{l \in D} l \beta_l(X_\infty)\right] = \mathbb{E}\left[\sum_{l \in D} \left[e(X_\infty + l) - e(X_\infty)\right] \beta_l(X_\infty)\right] = \pi Q e = 0.$$

In particular, looking at the first entry of F, we obtain the identity

$$\frac{\mathbb{E}[N_\infty - M_\infty]^+}{\mathbb{E}[M_\infty - N_\infty]^+} = \frac{c}{b}.$$

Indeed, Figure 4.3 shows how this ratio changes depending on the quotient c/b.

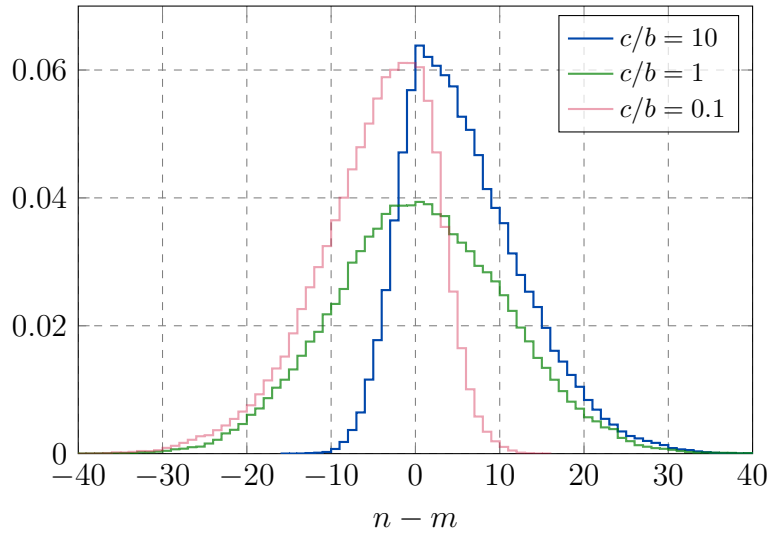


Figure 4.3: Histograms of $N_\infty - M_\infty$ for different ratios between b and c; the parameters of the simulations are $\lambda = 1000$ and $\mu = 1$.

This means that, in the steady-state, the ratio between the mean number of queued jobs and idle servers is determined by the lags b and c. These lags are inherent to the system and thus not under our control, which means that we cannot trade off queueing delay and over-provisioning at will. However, we will see in the next section that the latter can be achieved if we introduce some modifications in the provisioning rule. For simplicity we will assume $b = c$ in the sequel.

4.3 Controlling for zero queue length

We now plan to modify the provisioning rule that we have described above to manage the tradeoff between queue length and over-provisioning; now the question arises as to which of the two penalties is more troublesome from a practical perspective. According to recent literature on the subject [34], the entity in control of the server dynamics, in cloud-based systems and data centers, is a dispatcher which may not have enough local storage, or would rather avoid the overhead of holding jobs. Because of this, we will aim at the almost complete elimination of queueing.

The fluid dynamics of the system that we studied in the last section have a global attractor at the point (ρ, ρ). The equation $\dot{n} = \lambda - \mu \min(m, n)$, that appears in the dynamics (4.1), is inherent in the central queue scheme that we have adopted, and in particular independent of the provisioning rule that we choose.

Hence, a change in that rule will not move the system's equilibria away from the set $\{(m,n) \in [0,\infty)^2 : \min(m,n) = \rho\}$ that is depicted in Figure 4.4.

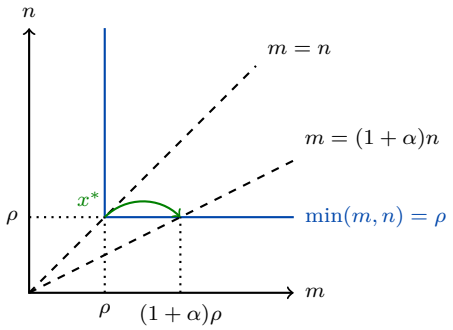


Figure 4.4: Feasible equilibria and shift of the equilibrium point of the dynamics (4.1).

In the last section we saw that, in the steady-state, the system hovers around the equilibrium of the fluid dynamics. Therefore, a strategy that could lead to the elimination of queueing is to move the equilibrium point of the dynamics (4.1) into the set $\{(m,n) \in [0,+\infty)^2 : m > n\}$, where the number of servers exceeds the number of jobs. To this end, fix some $\alpha \in (0,1)$ and consider the modification that Figure 4.5 introduces to the transitions diagram that appeared in Figure 4.2; we defer for now the discussion on implementation issues.

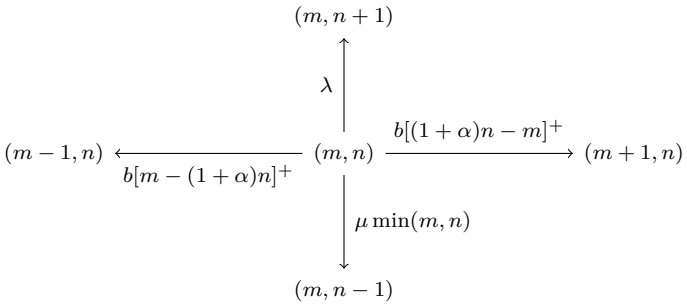


Figure 4.5: Modification of the Markovian model of Section 4.2, aiming to avoid queueing.

As it is illustrated in Figure 4.4, this modification is intended to shift the equilibrium of the dynamics (4.1) to the right, into the set $\{(m,n) \in [0,+\infty)^2 : m > n\}$. We will now perform a large scale analysis of the resulting system, in order to asses the effect of this change. To this purpose, we consider processes X, $\hat{X}_k$ and X_k analogous to those of Section 4.2; note that $\{X_k\}_{k\geq 1}$ is still a density dependent family, in this case generated by the maps

$$\beta_l(m,n) = \begin{cases} b[(1+\alpha)n - m]^+ & \text{if } l = (1,0), \\ b[m - (1+\alpha)n]^+ & \text{if } l = -(1,0), \\ \lambda & \text{if } l = (0,1), \\ \mu \min(m,n) & \text{if } l = -(0,1). \end{cases}$$

The drift of this family is still Lipschitz, and differentiable in the whole first quadrant except for the line $m = n$, indeed

$$F(m,n) = \begin{bmatrix} b[(1+\alpha)n - m] \\ \lambda - \mu \min(m,n) \end{bmatrix};$$

differentiability at the line $m = (1+\alpha)n$ results from the assumption that $b = c$, which simplifies computations significantly but does not modify the essence of the problem that we are discussing.

As in the previous section, we begin by computing the fluid limit of the family using Theorem 2.2.5. In this case, the macroscopic behavior of the Markov chains X_k is governed by the ODE

$$\begin{aligned} \dot{m} &= b[(1+\alpha)n - m], \\ \dot{n} &= \lambda - \mu \min(m,n). \end{aligned} \tag{4.2}$$

Proposition 4.3.1. The dynamics (4.2) have a unique equilibrium point x^*, with coordinates $m^* = (1+\alpha)\rho$ and $n^* = \rho$, and this equilibrium is a global attractor. Furthermore, it is even possible to find a common quadratic Lyapunov function for the dynamics (4.2).

Proof. Consider the matrices

$$A_1 = \begin{bmatrix} -b & (1+\alpha)b \\ -\mu & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} -b & (1+\alpha)b \\ 0 & -\mu \end{bmatrix},$$

the Jacobians of F in $\{(m,n) \in [0,\infty)^2 : m < n\}$ and $\{(m,n) \in [0,\infty)^2 : m > n\}$, respectively. We claim that there exists a positive definite symmetric matrix

$$P = \begin{bmatrix} 1 & q \\ q & r \end{bmatrix}$$

such that $A_i^T P + P A_i$ is negative definite for $i \in \{1,2\}$. If we let T_i and D_i denote the trace and determinant of the matrices $A_i^T P + P A_i$, then we must find $q, r \in \mathbb{R}$ such that P is positive definite and the following inequalities hold:

$$\begin{aligned} &T_1(q,r) = 2[(1+\alpha)b - \mu]q - 2b < 0, \\ &D_1(q,r) = -4(1+\alpha)b(b+\mu q)q - [(1+\alpha-q)b - \mu r]^2 > 0, \\ &T_2(q,r) = 2(1+\alpha)bq - 2\mu r - 2b < 0 \quad \text{and} \\ &D_2(q,r) = 4b[\mu r - (1+\alpha)bq] - [(1+\alpha-q)b - \mu q]^2 > 0. \end{aligned}$$

The set $\{(q,r) \in \mathbb{R}^2 : D_1(q,r) > 0\}$ is the interior of an ellipse that is located inside the strip $\{(q,r) \in \mathbb{R}^2 : -b/\mu < q < 0\}$ and is tangent to the line $q = 0$ at the point $(0, (1+\alpha)b/\mu)$. Also, $\{(q,r) \in \mathbb{R}^2 : D_2(q,r) > 0\}$ is the open set above the graph of a parabola that contains the point $(0, (1+\alpha)^2 b/(4\mu))$. These two sets are illustrated in Figure 4.6.

It is clear that the two sets intersect, because $(1+\alpha)b/\mu > (1+\alpha)^2 b/(4\mu)$ for all $\alpha \in (0,1)$. Moreover, there exists $\delta > 0$ such that $(-\varepsilon, (1+\alpha)b/\mu)$ lies in the

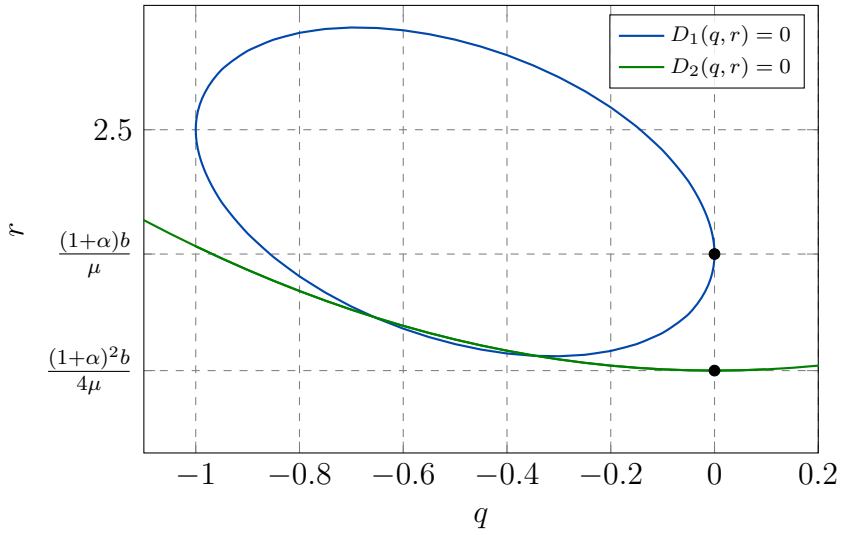


Figure 4.6: The conics $D_1(q,r)=0$ and $D_2(q,r)=0$ for $\mu=1$, $b=1$ and $\alpha=1/2$.

intersection for all $\varepsilon \in (0,\delta)$. Consequently, since

$$\lim_{\varepsilon\to 0} T_1\left(-\varepsilon, \frac{(1+\alpha)b}{\mu}\right) = -2b \quad \text{and} \quad \lim_{\varepsilon\to 0} T_2\left(-\varepsilon, \frac{(1+\alpha)b}{\mu}\right) = -2(2+\alpha)b,$$

there exists $\varepsilon > 0$ such that $q = -\varepsilon$ and $r = (1+\alpha)b/\mu$ are as desired. □

The interpretation of the last proposition is that the processes X_k are attracted towards the equilibrium point x^* as $t \to +\infty$. On the one hand, we see that the number of jobs still operates around ρ, as in the system of Section 4.2; this is a hard lower bound for the mean number of jobs, achieved by the ideal infinite-server queue. On the other hand, we are now accepting an over-provisioning of $\alpha\rho$ servers, and this will help us reduce the mean queue length at the dispatcher.

Another consequence of Proposition 4.3.1 is the following.

Corollary 4.3.2. The Markov chain X given by the transitions diagram that appears in Figure 4.5 is positive recurrent for all $\lambda, \mu, b > 0$ and $\alpha \in (0,1)$. In particular, the Markov chains $\hat{X}_k$ and X_k are positive recurrent for all $k \geq 1$.

Proof. The proof is as in Section 4.2, defining a Foster-Lyapunov function from the matrix P that we computed in Proposition 4.3.1, namely $V(y) = (y-x^*)^T P(y-x^*)$.

As in Section 4.2, if we let $D = \{\pm(1,0), \pm(0,1)\}$, then

$$QV(y) = 2(y-x^*)^T PF(y) + \sum_{l\in D} l^T Pl\beta_l(y) \quad \forall\, y \in \mathbb{R}^2.$$

Also, for all $y \in \{(m,n) \in [0,+\infty)^2 : m \geq n\}$ we have $F(y) = A_2(y-x^*)$ and thus

$$2(y-x^*)^T PF(y) = (y-x^*)^T (A_2^T P + PA_2)(y-x^*).$$

The only difference with Section 4.2 is that for $y \in \{(m,n) \in [0,+\infty)^2 : m < n\}$ we

have $F(y) = A_1(y - x^*) - \alpha\lambda[0\ 1]^T$, and hence

$$2(y-x^*)^T P F(y) = (y-x^*)^T (A_1^T P + P A_1)(y-x^*) - 2\alpha\lambda (y-x^*)^T P \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

However, $QV(y) \to -\infty$ as $y \to \infty$ still, and clearly $V(y) \to +\infty$ as $y \to \infty$. Therefore, the result follows from the Foster-Lyapunov criteria for positve recurrence that we stated in the previous section. □

Understanding how the system behaves near the equilibrium x^* warrants taking a closer look. To this end, we may use Theorem 3.4.4 to compute a central limit theorem around the point x^*; note that in this case Assumption 3.4.1 is automatic because F is differentiable at x^*. Consider then the matrices

$$A = A_2 = \begin{bmatrix} -b & (1+\alpha)b \\ 0 & -\mu \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2\lambda} \end{bmatrix}.$$

The first of these is the drift's Jacobian matrix at x^*, and the second is the dispersion coefficient of the SDE that appears in the statement of Theorem 3.4.4. In addition, if we let W be a bidimensional Wiener process, then this theorem tells us that, under suitable hypothesis on the initial conditions, the processes $Z_k = \sqrt{k}(X_k - x^*)$ converge weakly in $D_{\mathbb{R}^2}[0,\infty)$, as $k \to \infty$, to a process Z that solves

$$dZ_t = AZ_t dt + B dW_t.$$

This is a linear SDE where the drift coefficient A is a stable matrix: its eigenvalues have negative real parts. Therefore, using the arguments of Subsection 2.4.1 we may prove that this SDE is exponentially ergodic and that the invariant distribution Z_∞ is a bivariate Normal; in this case we have to resort to Remark D.3.5 because BB^T is singular. Furthermore, we know that Z_∞ has mean zero and its covariance matrix Σ_∞ is given by the Lyapunov equation

$$A\Sigma_\infty + \Sigma_\infty A^T + BB^T = 0.$$

The solution to this equation may be written in terms of the traffic intensity ρ, the fraction of over-provisioning α and the ratio $\eta = \mu/b$ between the mean server creation lag and the mean service time, specifically

$$\Sigma_\infty = \rho \frac{1+\alpha}{1+\eta} \begin{bmatrix} 1+\alpha & 1 \\ 1 & \frac{1+\eta}{1+\alpha} \end{bmatrix}. \tag{4.3}$$

The strong law of large numbers and the central limit theorem that we have computed suggest that $\hat{X}_k(\infty)$ is approximately $kx^* + \sqrt{k}Z_\infty$, in the steady-state and when k is large enough. We corroborate this numerically in Figure 4.7, by plotting a phase diagram of the system $\hat{X}_k$ and a level set of the Gaussian density corresponding to the random vector $kx^* + \sqrt{k}Z_\infty$.

Note that the system $\hat{X}_k$ receives jobs at rate $k\lambda$, its steady-state mean is located at the point $kx^* = (k\rho, k\rho)$ and the covariance of $\sqrt{k}Z_\infty$ is as in equation (4.3) but replacing ρ by the traffic intensity $k\rho$ that the system $\hat{X}_k$ faces. Hence, we

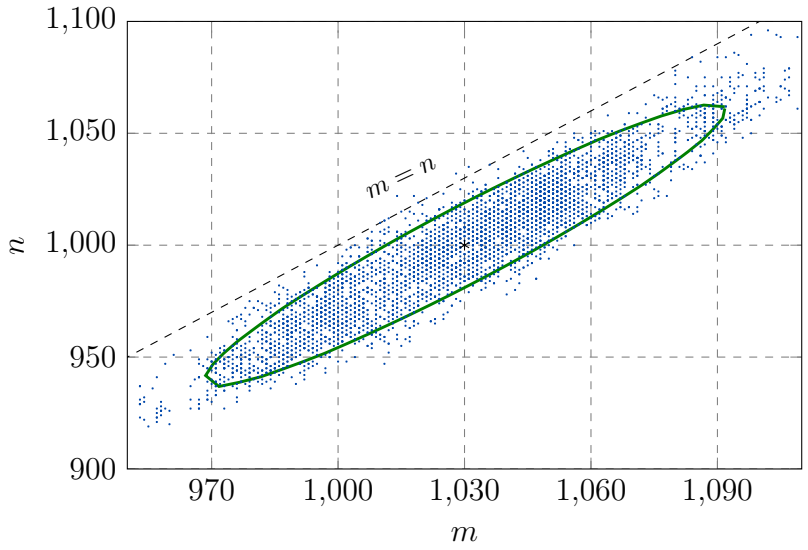


Figure 4.7: Level set of $(y-x^*)^T\Sigma_\infty^{-1}(y-x^*)$ and the states visited by a sample path of $\hat{X}_k$; the parameters of the simulation are $\lambda = 100$, $\mu = 1$, $b = 10$, $\alpha = 3\%$ and $k = 10$.

could rephrase the statement of the last paragraph saying that X_∞ is approximately $x^* + Z_\infty$ when λ is large enough, incorporating the scaling in the estimate. In particular, this suggests to approximate the difference $M_\infty - N_\infty$, between the number of servers and jobs, using a Normal random variable $N(\alpha\rho, \sigma^2)$ with variance

$$\sigma^2 = \begin{bmatrix} 1 & -1 \end{bmatrix} \Sigma_\infty \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{\alpha^2 + \eta}{1+\eta}\rho.$$

Using this approximation we may compute

$$\mathbb{P}(M_\infty - N_\infty \in [\alpha\rho - c\sigma, \alpha\rho + c\sigma]) = \mathbb{P}\left(\frac{M_\infty - N_\infty - \alpha\rho}{\sigma} \in [-c, c]\right) \approx 2\Phi(c) - 1,$$

where Φ is the cumulative distribution function of the standard Normal. This estimate may be used to design the system to avoid queueing with high probability.

Indeed, we may choose c so that the right-hand side of the above equation is close to one, and then compute α so that $\alpha\rho - \sigma c > 0$, or equivalently

$$\frac{1}{\rho(1+\eta)} + \frac{\eta}{\rho\alpha^2(1+\eta)} < \frac{1}{c^2}. \tag{4.4}$$

For instance, in the simulation of Figure 4.8 we computed α so that the above condition held for $c = 2$; in this case we have $2\Phi(c) - 1 > 0.95$. This simulation shows how the difference between the number of jobs and servers stays within the confidence interval $[\alpha\rho - 2\sigma, \alpha\rho + 2\sigma]$ with high probability, thus avoiding queueing.

We may also use our Gaussian approximation to estimate the mean queue length in the steady-state. Letting φ be the density of the standard Normal we have

$$\mathbb{E}[N_\infty - M_\infty]^+ \cong \sigma\varphi\left(\frac{\alpha\rho}{\sigma}\right) - \alpha\rho\Phi\left(\frac{-\alpha\rho}{\sigma}\right).$$

This function of α is plotted in Figure 4.9 for different traffic intensities ρ; there

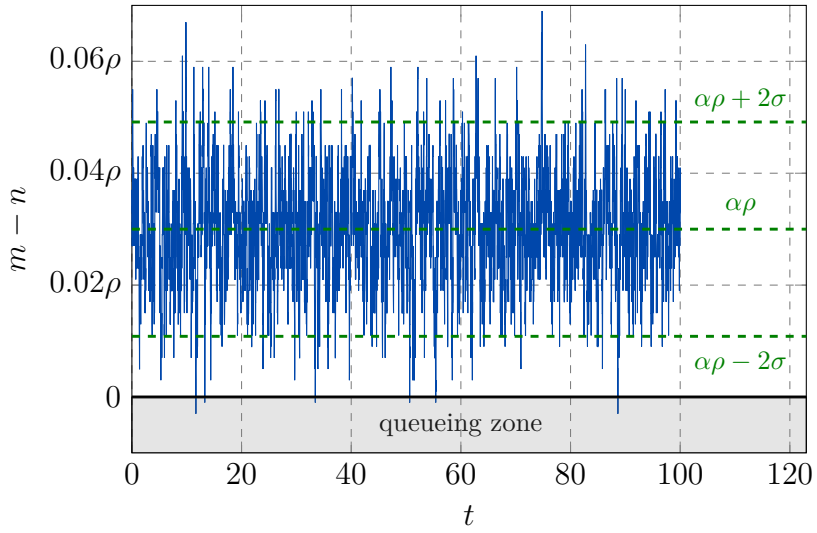


Figure 4.8: Simulation of the system X showing the over-provisioning level and the avoidance of the queueing zone; the parameters of the simulation are $\lambda = 1000$, $\eta = 0.1$ and $\alpha = 3\%$.

we see that the mean queue length approaches zero rapidly as α increases, which is reasonable since $\alpha\rho$ is the system's over-provisioning level. For instance, when the traffic intensity is $\rho = 1000$, a 2% over-provisioning yields a mean queue length of order two, and a 5% over-provisioning results in nearly zero queue at the dispatcher.

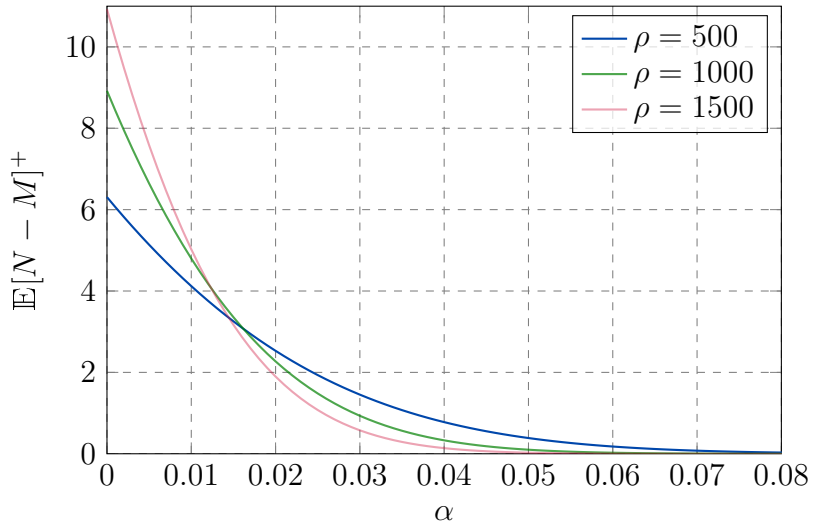


Figure 4.9: Steady-state estimate of the mean queue length at the dispatcher for $\eta = 1$.

For the plots in Figure 4.9 we assumed $\eta = 1$, which means that the mean creation lag of servers is equal to the mean service time of jobs. Clearly the performance is better when η is smaller than one as Figure 4.8 shows; there a 3% over-provisioning is enough to eliminate queueing almost completely when $\rho = 1000$. On the contrary, the performance declines when η is greater than one; this is captured by our model since an increase in η causes an increase in σ^2. Still, for reasonable values of η a moderate amount of over-provisioning yields almost zero queue at the dispatcher.

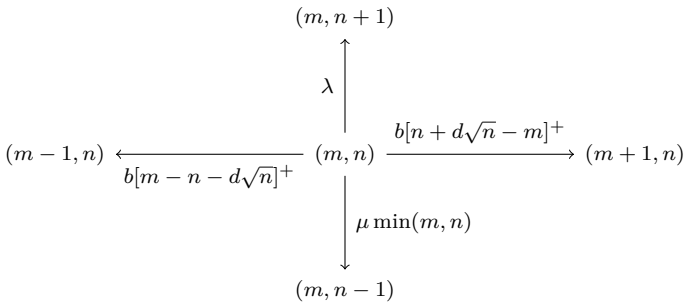


Figure 4.10: Automatic rule for adjusting the over-provisioning level to the uncertain load.

Besides, we see in Figure 4.9 that the queue length decays more sharply with α for higher values of ρ. This suggests that, rather than selecting a fixed fraction of over-provisioning, we could try to adapt the over-provisioning level to the uncertain traffic intensity ρ. We return to equation (4.4) with this in mind, and see that in order to satisfy this criteria one must let $\alpha = O(1/\sqrt{\rho})$. Equivalently, the minimum over-provisioning that we need to avoid queuing is $\alpha\rho = O(\sqrt{\rho})$ as in the Halfin-Whitt regime of the many-server queue; recall the squre root staffing rule that we explained in Remark 3.1.1. In the next section we propose a method that self-adjusts the number of idle servers to this level.

4.4 Automatic control of the over-provisioning

We now focus on an automatic rule, independent of the traffic intensity ρ, with the aim of achieving the desired over-provisioning level of $O(\sqrt{\rho})$ servers. The main idea is to replace the constant α in the transitions diagram of Figure 4.5 by a function of the form $\alpha(n) = d/\sqrt{n}$, approximating ρ by its instantaneous estimate, the current occupation level n. The chain X that results from this modification is shown in Figure 4.10.

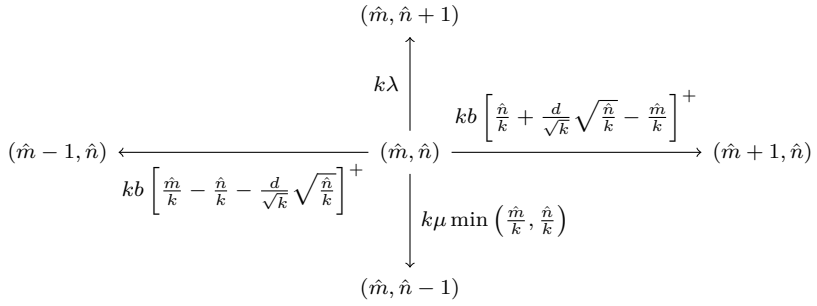


Figure 4.11: Transition rates of the processes $\hat{X}_k$.

As before we consider the systems $\hat{X}_k$ that face arrival rates of $k\lambda$ jobs per second; recall that these systems are described by the chain of Figure 4.10 with λ replaced by $k\lambda$. The normalization $X_k = \hat{X}_k/k$ yields once more a density dependent family,

but in this case the intensities are given by maps of the form $\beta_l^k = \gamma_l + \delta_l^k$, where the perturbation terms δ_l^k are non-zero. To compute these maps we first write the intensities of $\hat{X}_k$ as in Figure 4.11, and then we perform the change of variables $m = \hat{m}/k$ and $n = \hat{n}/k$. In this way we find expressions for the maps $\beta_l^k(m,n)$, which we may write in terms of:

$$\gamma_l(m,n) = \begin{cases} b[n-m]^+ & \text{if } l = (1,0), \\ b[m-n]^+ & \text{if } l = -(1,0), \\ \lambda & \text{if } l = (0,1), \\ \mu \min(m,n) & \text{if } l = -(0,1). \end{cases} \quad \text{and}$$

$$\delta_l^k(m,n) = \begin{cases} b\left[n + \frac{d}{\sqrt{k}}\sqrt{n} - m\right]^+ - b[n-m]^+ & \text{if } l = (1,0), \\ b\left[m - n - \frac{d}{\sqrt{k}}\sqrt{n}\right]^+ - b[m-n]^+ & \text{if } l = -(1,0), \\ 0 & \text{if } l = (0,1), \\ 0 & \text{if } l = -(0,1). \end{cases}$$

Furthermore, we see from the above definitions that the drift and perturbing drifts are given, respectively, by the expressions

$$F(m,n) = \begin{bmatrix} b(n-m) \\ \lambda - \mu \min(m,n) \end{bmatrix} \quad \text{and} \quad G_k(m,n) = \frac{bd}{\sqrt{k}} \begin{bmatrix} \sqrt{n} \\ 0 \end{bmatrix}.$$

The drift F is Lipschitz. Furhtermore, it is easy to check that the other hypothesis of Theorem 2.2.5 hold as well. Consequently, the fluid limit of the processes X_k is given by the dynamics

$$\begin{aligned} \dot{m} &= b(n-m), \\ \dot{n} &= \lambda - \mu \min(m,n). \end{aligned} \tag{4.5}$$

Note that this is the same ODE of Section 4.2, the only difference is that we are now assuming that $b = c$. In particular, $x^* = (\rho, \rho)$ is a global attractor of the dynamics, and thus the number of jobs and servers both operate around ρ in the fluid scale, which corresponds to zero over-provisioning. The rationale is that the number of idle servers in the system, which operates around $\sqrt{\rho}$, is negligible in this macroscopic scale when $\rho \to +\infty$. Hence, in order to see how the system counteracts the queuing delay, we need to look into the diffusion scale.

To this purpose, we will make use of Theorem 3.4.4, but first we must verify its hypothesis. To begin, we note that the field ∂F of Assumption 3.4.1 may be defined as in Section 4.2. Specifically, let $\nu = [-1\ 1]^T$ and consider the matrices

$$A_1 = \begin{bmatrix} -b & b \\ -\mu & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} -b & b \\ 0 & -\mu \end{bmatrix},$$

the Jacobians of F in $\{(m,n) \in [0,+\infty)^2 : m < n\}$ and $\{(m,n) \in [0,+\infty)^2 : m > n\}$, respectively; we may then let

$$\partial F(y) = A_1 y \mathbb{1}_{\langle y,\nu \rangle \geq 0} + A_2 y \mathbb{1}_{\langle y,\nu \rangle < 0}.$$

It is furthermore easy to check that Assumption 3.2.1 holds, and Assumption 3.4.3

holds as well if we let

$$G(m,n) = bd \begin{bmatrix} \sqrt{n} \\ 0 \end{bmatrix}.$$

Consequently, we may indeed use Theorem 3.4.4, under the hypothesis on the initial conditions that are stated therein. According to this theorem, the processes $Z_k = \sqrt{k}(X_k - x^*)$ converge weakly in $D_{\mathbb{R}^2}[0,\infty)$ to a diffusion Z such that

$$dZ_t = [\partial F(Z_t) + G(x^*)]dt + BdW_t,$$

where W is a bidimensional Wiener process and B is given by

$$B = \begin{bmatrix} \sqrt{\gamma_{(1,0)}(x^*) + \gamma_{(-1,0)}(x^*)} & 0 \\ 0 & \sqrt{\gamma_{(0,1)}(x^*) + \gamma_{(0,-1)}(x^*)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2\lambda} \end{bmatrix}.$$

In order to see how the system's over-provisioning becomes apparent in the diffusion scale, it helps to write the dynamics of $Z = (U, V)$ in coordinates:

$$\begin{aligned} dU_t &= b\,(V_t - U_t + d\sqrt{\rho})\,dt, \\ dV_t &= -\mu \min(U_t, V_t)dt + \sqrt{2\lambda}dW_t. \end{aligned}$$

The offset $d\sqrt{\rho}$ in the first of these equations suggests that the system operates with $O(\sqrt{\rho})$ idle servers, and this is confirmed by the simulations of the following section.

Unfortunately, the switching in the drift coefficient precludes us from computing an invariant measure. However, in the sequel we will corroborate numerically that the estimates of the preceding section can be used to predict the performance of the system that we have described here. The difference between the current system and that of Section 4.3 is that we have replaced the constant over-provisioning fraction α by a function $\alpha(n) = d/\sqrt{n}$ that tracks $d/\sqrt{\rho}$. Hence, it is reasonable to expect the same behavior that we saw in Section 4.3 for $\alpha = d/\sqrt{\rho}$.

Remark 4.4.1. The most suitable value of the constant d depends on the ratio η between the mean server creation lag and the mean service time, and the criteria of equation (4.4) provides a practical rule for choosing d. Indeed, letting $\alpha = d/\sqrt{\rho}$ this equation becomes

$$\frac{1}{\rho(1+\eta)} + \frac{\eta}{d(1+\eta)} < \frac{1}{c^2},$$

where c is chosen in advance to ensure that queuing is avoided with high probability. In the above equation, the first term is negligible when the traffic intensity is high; we may then compute d in terms of c and η.

4.5 Implementation and further simulations

We begin this section providing an implementation of the provisioning rule that we studied in Section 4.3, which requires to determine how server creation and deletion requests have to be managed. For example, the transitions in Figure 4.5

that correspond to the creation of servers should take place at a rate that is b times the number of pending server requests.

In the boundary case $\alpha = 0$, which corresponds to the system of Section 4.2, we were able to provide an exact implementation of the Markov chain that appeared in Figure 4.1. Unfortunately, this will not be possible in the general case, because that would require to maintain non-integral numbers of pending server and shut down requests. For instance, in the chain of Figure 4.5 servers are created at rate $b[(1+\alpha)n - m]^+$ and the number $[(1+\alpha)n - m]^+$ is in general not an integer.

An approximate implementation is however possible. Let $r(x)$ denote the integer that lies closest to $x \in \mathbb{R}$. In the implementation that we propose, the dispatcher keeps track of the number $q(m,n) = (1+\alpha)n - m$, updating it whenever the number of servers or jobs changes. This variable is used to compute a target value for the number of server and shut down requests that should be pending; in the case of server requests the target value is $r([q(m,n)]^+)$, while in the case of shut down requests the target is $r([-q(m,n)]^+)$. The actual number of either of these requests is kept aligned with the corresponding target by issuing or withdrawing requests when deviations occur.

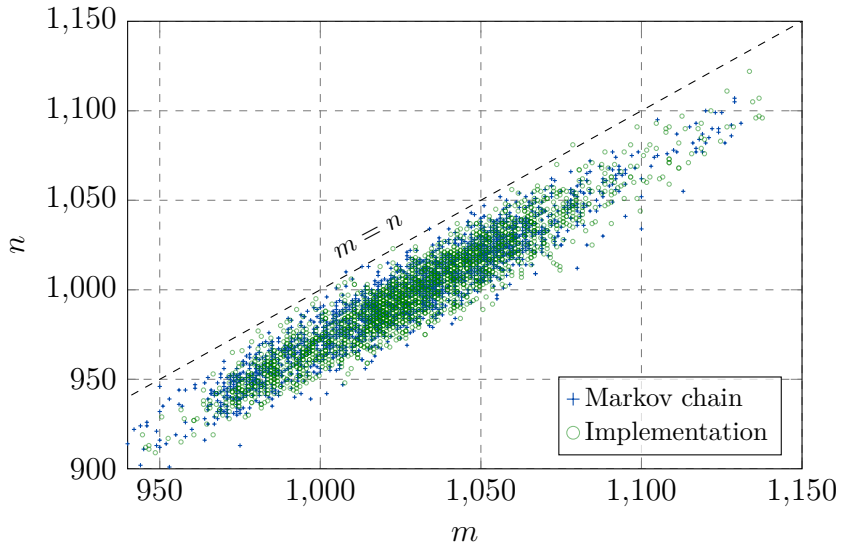


Figure 4.12: Simulation of the Markov chain of Figure 4.10 and the proposed implementation, the plot shows the states that each of these systems visited. The parameters of the simulation are $\lambda = 1000$, $\mu = 1$, $b = 10$ and $d = 1$.

A similar implementation is possible for the system of Section 4.4; the algorithm is as above but replacing the constant α with the function $\alpha(n) = d/\sqrt{n}$. This implementation is compared with the Markov chain of Figure 4.10 by means of the simulation that we plot in Figure 4.12; the similarity between the two phase diagrams suggests that the approximations that we have made are accurate.

Furthermore, Table 4.1 compares time averages of the relevant metrics with the Gaussian estimates of Section 4.3, and we only see minor differences; the estimates

	Time averages (Markov chain)	**Time averages (Implementation)**	**Estimates**
$\mathbb{E}[M]$	1031	1029	1032
$\mathbb{E}[N]$	999	997	1000
$\mathbb{V}[M-N]$	94	96	92
$\mathbb{E}[M-N]^+$	32	31	32
$\mathbb{E}[N-M]^+$	0.0007	0.0007	0.0012

Table 4.1: Data corresponding to simulations with $\lambda = 1000$, $\mu = 1$, $b = 10$ and $d = 1$. The first row corresponds to a simulation of the chain of Figure 4.10, the second row corresponds to the approximate implementation that we proposed and the third row shows the estimates that we computed in Section 4.3 for $\alpha = d/\sqrt{\rho}$.

are evaluated considering an over-provisioning fraction of $\alpha = d/\sqrt{\rho}$. As we commented at the end of Section 4.4, the only difference between the systems of sections 4.3 and 4.4 is that, in the second, the static over-provisioning fraction α is replaced by a dynamic over-provisioning fraction $\alpha(n) = d/\sqrt{n}$. However, since this function is designed to track $d/\sqrt{\rho}$, it is therefore reasonable to expect a behavior similar to that of Section 4.3 when $\alpha = d/\sqrt{\rho}$.

Chapter 5

Conclusions

In this thesis we discussed functional laws of large numbers and central limit theorems for density dependent families of continuous time Markov chains. These are natural generalizations of their analogs for random variables, in the sense that the law of large numbers yields a deterministic limit, whereas the central limit theorem produces a Gaussian limit; at least in its classical version. In the dynamic case the limits are governed, respectively, by an ODE and a SDE; a very elegant feature is that these arise from the Markovian dynamics of the family.

A wide variety of continuous parameter families of Markov chains arise naturally in applied probability and fall into the category of density dependent families; for instance in epidemics, chemistry and stochastic networks. The above limit theorems can be used to obtain useful quantitative estimates of the metrics that are relevant for the application; the quality of these approximations can always be judge by numerical comparisons. Besides these estimates, the methodology provides a valuable insight on the asymptotic behavior of the metrics and their relative orders of magnitude as the parameter of the family approaches infinity; moreover, it indicates the regions where the numerical approximations may fail.

Some of the density dependent families that arise in the stochastic analysis of networks do not fit the hypothesis of the classic central limit theorem due to Kurtz. The latter motivated us to extend this theorem in two directions: to contemplate small order perturbations in the intensities of the density dependent family and to consider non-differentiable drifts. Families with these characteristics had been studied in the literature before, but only in particular cases, and to our knowledge a general treatment of the problem had not been performed until this work.

The central limit theorem that we developed for families with a non-differentiable drift produces a limit that is governed by a SDE with switching in the drift coefficient. In general, the analysis of the corresponding diffusion goes beyond the state of the art techniques, and even proving its ergodicity is usually non-trivial; we note however that we did carry out the analysis in the unidimensional case. Understanding these diffusions is a very difficult and interesting problem that is connected to the study of elliptic equations for measures and PDEs.

The classical limit theorems for density dependent families, and the extensions that we developed, were used to study the dynamic right sizing of computing resources in large scale cloud environments and data centers; the goal was to design a provisioning rule capable of adjusting the active computing capacity to an uncertain workload. Since we opted for a central queue scheme, we were particularly interested in eliminating queueing, because storing a large amount jobs in a single queue can be problematic from a technological perspective.

With the latter in mind, we proposed a rule capable of eliminating queueing almost completely, at the expense of a small amount of over-provisioning: for a traffic intensity of ρ, the number of idle servers scales as $O(\sqrt{\rho})$ when the arrival rate of jobs approaches infinity. In other words, the number of active servers operates around $\rho + O(\sqrt{\rho})$, and in this sense our policy tracks the celebrated Halfin-Whitt regime in a automatic fashion. The analysis of this provisioning rule was carried out using the limit theorems that we developed throughout the thesis and by means of numerical simulations.

Appendix A

Weak convergence in Skorohod spaces

This appendix contains some basic definitions and results regarding the weak convergence of probability measures in Skorohod spaces. Its whole content is based on [2, Chapter 1, Chapter 3], where the reader may find complete proofs of the results that are stated below.

A.1 Weak convergence

We begin with some general facts regarding the weak convergence of probability measures in metric spaces. To this end, consider a metric space (E, ρ), let $\mathcal{B}(E)$ denote its Borel σ-algebra and define $C_b(E)$ to be the set of all continuous and bounded functions $f : E \longrightarrow \mathbb{R}$. The notation

$$Pf = \int_E f dP$$

will be used to denote the integral of a function $f \in C_b(E)$ with respect to a probability measure P on $\mathcal{B}(E)$. In the sequel P and P_n will always denote probability measures on $\mathcal{B}(E)$.

Definition A.1.1. The sequence of probability measures P_n converges weakly to P if $P_n f \to Pf$ as $n \to \infty$ for all $f \in C_b(E)$, and we denote this by writing $P_n \Rightarrow P$.

The uniqueness of the limit is given by the next theorem.

Theorem A.1.2. Two probability measures P and Q on $\mathcal{B}(E)$ coincide if and only if $Pf = Qf$ for all bounded and uniformly continuous $f : E \longrightarrow \mathbb{R}$.

A set $A \subset E$ is said to be a P-continuity set if $P(\partial A) = 0$; here ∂A denotes the boundary of A, which is a closed set, and hence belongs to $\mathcal{B}(E)$. We may now state the "portmanteau" theorem, which characterizes weak convergence.

Theorem A.1.3 (Portmanteau theorem)**.** The following statements are equivalent.

1. $P_n \Rightarrow P$ as $n \to \infty$.
2. $P_n f \to Pf$ as $n \to \infty$ for all bounded and uniformly continuous $f : E \longrightarrow \mathbb{R}$.
3. $\limsup_{n\to\infty} P_n(F) \leq P(F)$ for all closed sets $F \subset E$.
4. $\liminf_{n\to\infty} P_n(G) \geq P(G)$ for all open sets $G \subset E$.
5. $P_n(A) \to P(A)$ as $n \to \infty$ for all P-continuity sets $A \subset E$.

Let $h : E \longrightarrow F$ be a measurable map between metric spaces. For each probability measure P on $\mathcal{B}(E)$ this map induces another probability measure Ph^{-1} on $\mathcal{B}(F)$, which is given by $Ph^{-1}(A) = P(h^{-1}(A))$ for all $A \in \mathcal{B}(F)$. A straightforward, although very useful result, is the following.

Theorem A.1.4 (Continuous mapping theorem)**.** Consider a continuous mapping $h : E \longrightarrow F$ between two metric spaces. If $P_n \Rightarrow P$ in E as $n \to \infty$, then $P_n h^{-1} \Rightarrow Ph^{-1}$ in F as $n \to \infty$.

As a matter of fact we even have the following refinement.

Theorem A.1.5. Consider a map $h : E \longrightarrow F$ between metric spaces and let D_h be the set of its discontinuities. If $P_n \Rightarrow P$ in E as $n \to \infty$, and $P(D_h) = 0$, then $P_n h^{-1} \Rightarrow Ph^{-1}$ in F as $n \to \infty$.

A.1.1 Convergence in distribution

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A random element is just a measurable map $X : \Omega \longrightarrow E$, and each random element induces a probability measure on $\mathcal{B}(E)$, namely the measure $\mathbb{P}X^{-1}$. A sequence of random elements X_n converges in distribution to X if the corresponding measures $\mathbb{P}X_n^{-1}$ converge weakly to $\mathbb{P}X^{-1}$, or equivalently $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$ for all $f \in C_b(E)$. We will use the notation $X_n \Rightarrow X$ to denote convergence in distribution.

Definition A.1.6. The sequence of random elements X_n converges in probability to the constant $x \in E$ if $\mathbb{P}(\rho(X_n, x) < \varepsilon) \to 1$ as $n \to \infty$ for all $\varepsilon > 0$, and we denote this by writing $X_n \xrightarrow{\mathbb{P}} x$.

By the fourth item of the portmanteau theorem $X_n \xrightarrow{\mathbb{P}} x$ is equivalent to $X_n \Rightarrow x$; here x is being regarded as a constant random element whose corresponding probability measure is the unit mass at x.

Consider two metric spaces (E_1, ρ_1) and (E_2, ρ_2), the product space $E_1 \times E_2$ may be regarded as a metric space with the product topology; for instance, this is the case if we endow $E_1 \times E_2$ with the metric

$$\varrho((x_1, x_2), (y_1, y_2)) = \max(\rho_1(x_1, y_1), \rho_2(x_2, y_2)).$$

In what follows, the product spaces $E_1 \times E_2$ are always endowed with some metric that generates the product topology.

Theorem A.1.7. Let (X_n, Y_n) be random elements in $E \times E$. If $X_n \Rightarrow X$ in E, and $\rho(X_n, Y_n) \xrightarrow{\mathbb{P}} 0$ in $\mathbb{R}$, as $n \to \infty$, then $Y_n \Rightarrow X$ in E as $n \to \infty$.

Suppose that (X, Y) is a random element in $E_1 \times E_2$, then X and Y are random elements in E_1 and E_2, respectively; because the natural projections are continuous. Furthermore, the converse is true if E_1 and E_2 are separable. The next propositions do not appear in [2], and thus we prove them, even though they are a almost straightforward consequence of the last theorem.

Proposition A.1.8. Consider two separable metric spaces E_1 and E_2. Suppose that $X_n \Rightarrow X$ in E_1, and $Y_n \xrightarrow{\mathbb{P}} y$ in E_2, as $n \to \infty$; where $y \in E_2$ is a constant. Then $(X_n, Y_n) \Rightarrow (X, y)$ in $E_1 \times E_2$ as $n \to \infty$.

Proof. Choose some $f \in C_b(E_1 \times E_2)$. It is clear that $x \mapsto f(x, y)$ is a continuous and bounded map on E_1, and therefore $\mathbb{E}[f(X_n, y)] \to \mathbb{E}[f(X, y)]$ as $n \to \infty$. As a result $(X_n, y) \Rightarrow (X, y)$ in $E_1 \times E_2$ as $n \to \infty$.

Suppose that $E_1 \times E_2$ is endowed with the metric

$$\rho((x_1, x_2), (y_1, y_2)) = \max(\rho_1(x_1, y_1), \rho_2(x_2, y_2)),$$

which generates the product topology; here ρ_1 and ρ_2 are, respectively, the metrics of E_1 and E_2. The product space $E = E_1 \times E_2$ is separable, this implies that $((X_n, Y_n), (X_n, y))$ is a random element in $E \times E$. Hence, the observation that

$$\mathbb{P}\left(\rho((X_n, Y_n), (X_n, y)) < \varepsilon\right) = \mathbb{P}\left(\rho_2(Y_n, y) < \varepsilon\right) \to 1 \quad \text{as } n \to \infty \quad \forall\, \varepsilon > 0$$

completes the proof by Theorem A.1.7. □

A random element $X : \Omega \longrightarrow \mathbb{R}$ is called a random variable.

Proposition A.1.9. Let E be a separable Banach space. If $X_n \Rightarrow X$ in E, and $Y_n \xrightarrow{\mathbb{P}} 0$ in $\mathbb{R}$, as $n \to \infty$, then $X_n Y_n \xrightarrow{\mathbb{P}} 0$ in E as $n \to \infty$.

Proof. By Proposition A.1.8 we know that $(X_n, Y_n) \Rightarrow (X, 0)$ in $E \times \mathbb{R}$ as $n \to \infty$. The result now follows from the continuous mapping theorem; recall that convergence in probability and convergence in distribution to a constant random element are the same thing. □

A.1.2 The Prohorov theorem

The following notion of relative compactness is very useful for proving the weak convergence of probability measures.

Definition A.1.10. A family of probability measures Π on $\mathcal{B}(E)$ is said to be relatively compact if each sequence contained in Π has a subsequence that converges weakly to some probability measure on $\mathcal{B}(E)$.

The strategy is to use relative compactness together with the following result.

Proposition A.1.11. $P_n \Rightarrow P$ as $n \to \infty$ if and only if each subsequence $(P_{n_m})_{m\geq 1}$ contains a further subsequence $(P_{n_{m_k}})_{k\geq 1}$ such that $P_{n_{m_k}} \Rightarrow P$ as $k \to \infty$.

When the sequence $(P_n)_{n\geq 1}$ is relatively compact, we already now that each subsequence $(P_{n_m})_{m\geq 1}$ has a further subsequence $(P_{n_{m_k}})_{k\geq 1}$ that converges weakly to some probability measure Q on $\mathcal{B}(E)$. Hence, if we want to show that $P_n \Rightarrow P$ as $n \to \infty$, we only need to prove that Q is always equal to P. The point is that we only have to deal with the problem of characterizing the limit, and we may dodge the problem of proving its existence.

If we want to embrace this approach, then we need effective means of proving relative compactness. It is usually easier to prove tightness, which is defined below.

Definition A.1.12. A family of probability measures Π on $\mathcal{B}(E)$ is tight if for each $\varepsilon > 0$ there exists a compact set $K \subset E$ such that $P(K) > 1 - \varepsilon$ for all $P \in \Pi$.

The relation between relative compactness and tightness is given by the following theorem, which is due to Prohorov.

Theorem A.1.13 (Prohorov theorem)**.** Let Π be a family of probability measures on $\mathcal{B}(E)$. If Π is tight, then Π is relatively compact. Moreover, if E is separable and Π is relatively compact, then Π is tight as well.

A.2 The space $D_{\mathbb{R}^d}[0,T]$

Throughout this section we are going to consider a fixed interval $[0,T]$, and for each $x : [0,T] \longrightarrow \mathbb{R}^d$ we will let

$$x(t^-) = \lim_{s\to t^-} x(s) \quad \text{and} \quad x(t^+) = \lim_{s\to t^+} x(s),$$

whenever the limits are defined and exist. We say that x has left limits if the first of these limits exists at all $t \in (0,T]$, and we say that x is right continuous if the second of these limits exists, and moreover $x(t^+) = x(t)$, at all $t \in [0,T)$.

Definition A.2.1. The Skorohod space $D_{\mathbb{R}^d}[0,T]$ consists of all right continuous functions $x : [0,T] \longrightarrow \mathbb{R}^d$ with left limits, which are usually called càdlàg functions.

For each $x : [0,T] \longrightarrow \mathbb{R}^d$ we have the following continuity moduli.

$$w_x(S) = \sup_{s,t\in S} ||x(t) - x(s)|| \quad \forall\ S \subset [0,T] \quad \text{and}$$

$$w_x(\delta) = \sup_{|t-s|\leq\delta} ||x(t) - x(s)|| \quad \forall\ \delta > 0.$$

The analog of the uniform continuity of continuous functions with a compact domain is given in the next proposition.

Proposition A.2.2. For each $x \in D_{\mathbb{R}^d}[0,T]$ and each $\varepsilon > 0$, there exists a partition $0 = t_0 < \cdots < t_n = T$ such that $w_x[t_{i-1}, t_i) < \varepsilon$ for all $i = 1, \ldots, n$.

The preceding proposition implies that, given $\varepsilon > 0$, each càdlàg function x has finitely many discontinuity points t where $||x(t) - x(t^-)|| \geq \varepsilon$. As a result, càdlàg functions have at most countably many discontinuities. Another consequence of the last proposition is the following.

Corollary A.2.3. If $x \in D_{\mathbb{R}^d}[0,T]$, then x has a bounded range.

The moduli of continuity $w_x(\delta)$ are adequate for characterizing continuous functions; for instance, $w_x(\delta) \to 0$ as $\delta \to 0$ if and only if x is continuous. In order to define the analog moduli for càdlàg functions, let us consider the collection Π_δ of all partitions $\pi = \{0 = t_0 < \cdots < t_{n_\pi} = T\}$ such that $t_i - t_{i-1} > \delta$ for all $i = 1, \ldots, n_\pi$. Now we may define the moduli

$$w'_x(\delta) = \inf_{\pi \in \Pi_\delta} \max_{1 \leq i \leq n_\pi} w_x[t_{i-1}, t_i) \quad \forall\ \delta \in (0,T).$$

Proposition A.2.4. A map $x : [0,T] \longrightarrow \mathbb{R}^d$ belongs to $D_{\mathbb{R}^d}[0,T]$ if and only if

$$\lim_{\delta \to 0} w'_x(\delta) = 0.$$

The moduli $w_x(\delta)$ and $w'_x(\delta)$ are essentially the same for continuous functions. Indeed, if for each càdlàg x we let $j(x) = \max\{||x(t) - x(t^-)|| : t \in [0,T]\}$, then

$$w'_x(\delta) \leq w_x(2\delta) \leq 2w'_x(2\delta) + j(x);$$

note that the maximum in the definition of $j(x)$ is attained by Proposition A.2.2.

A.2.1 The Skorohod topology

Suppose just for a moment that $d = 1$ and consider two continuous functions $x, y : [0,T] \longrightarrow \mathbb{R}$. These functions are said to be near in the uniform topology if the graph of x can be carried out onto the graph of y by means of an uniformly small perturbation of the ordinates, keeping the abscissas fixed. In the Skorohod topology we will allow a uniformly small deformation of the time scale as well. The deformation in the time scale will be given by a continuous and increasing bijection $\lambda : [0,T] \longrightarrow [0,T]$; the set of all these bijections will be denoted Λ.

Definition A.2.5. For each $x, y \in D_{\mathbb{R}^d}[0,T]$ we define

$$d(x,y) = \inf_{\lambda \in \Lambda} \max \left\{ \sup_{t \in [0,T]} |\lambda(t) - t|,\ \sup_{t \in [0,T]} ||x(\lambda(t)) - y(t)|| \right\}.$$

Equivalently, $d(x,y)$ is the infimum of those $\varepsilon > 0$ for which there exists $\lambda \in \Lambda$ such that the following hold.

$$\sup_{t \in [0,T]} |\lambda(t) - t| < \varepsilon \quad \text{and} \quad \sup_{t \in [0,T]} ||x(\lambda(t)) - y(t)|| < \varepsilon.$$

Note that the boundedness of càdlàg functions implies that d is always finite, and it is easy to check that d is a metric; the corresponding topology is called the Skorohod topology. Convergence in the Skorohod topology implies pointwise convergence at continuity points. Specifically, if $d(x_n, x) \to 0$ as $n \to \infty$ and x is continuous at t, then $x_n(t) \to x(t)$ as $n \to \infty$. Moreover, the Skorohod topology relativized to the subspace $C_{\mathbb{R}^d}[0,T]$, of continuous functions, coincides with the uniform topology.

The metric space $(D_{\mathbb{R}^d}[0,T], d)$ has the disadvantage of not being complete. However, it is possible to define a metric d_0, equivalent to d in the sense that it generates the same topology, and such that $(D_{\mathbb{R}^d}[0,T], d_0)$ is complete. To this end, we define for each non-decreasing $\lambda : [0,T] \longrightarrow [0,T]$ the quantity

$$||\lambda|| = \sup_{s<t} \left| \log \left(\frac{\lambda(t) - \lambda(s)}{t-s} \right) \right|.$$

When $||\lambda||$ is finite the slopes of the chords of λ are bounded away from zero and infinity. This implies that λ is continuous and strictly increasing, thus $\lambda \in \Lambda$; note however that λ may belong to Λ and still $||\lambda||$ may not be finite.

Definition A.2.6. For each $x, y \in D_{\mathbb{R}^d}[0,T]$ we define

$$d_0(x,y) = \inf_{\lambda \in \Lambda} \max \left\{ ||\lambda||, \sup_{t \in [0,T]} ||x(\lambda(t)) - y(t)|| \right\}.$$

Equivalently, $d_0(x,y)$ is the infimum of those $\varepsilon > 0$ for which there exists $\lambda \in \Lambda$ such that the following conditions hold.

$$||\lambda|| < \varepsilon \quad \text{and} \quad \sup_{t \in [0,T]} ||x(\lambda(t)) - y(t)|| < \varepsilon.$$

As we commented d_0 is a metric. Furthermore, we have the next theorem.

Theorem A.2.7. The metrics d and d_0 are equivalent, and $(D_{\mathbb{R}^d}[0,T], d_0)$ is a complete and separable metric space.

Recall that a subspace Y of a topological space X is said to be relatively compact if its closure is a compact set; in the case of metric spaces it is equivalent to say that Y is relatively compact if every sequence in Y has a converging subsequence.

Theorem A.2.8. A set $A \subset D_{\mathbb{R}^d}[0,T]$ is relatively compact if and only if

$$\sup_{x \in A} \sup_{t \in [0,T]} ||x(t)|| < \infty \quad \text{and} \quad \lim_{\delta \to 0} \sup_{x \in A} w'_x(\delta) = 0.$$

The last theorem is the analog of the Arzelá-Ascoli theorem, but for the Skorohod topology instead of the uniform topology, which is used with continuous functions.

A.2.2 Finite-dimensional sets

Given $0 \leq t_1 < \cdots < t_n \leq T$ we define $\pi_{t_1,\ldots,t_k} : D_{\mathbb{R}^d}[0,T] \longrightarrow \mathbb{R}^k$ such that

$$\pi_{t_1,\ldots,t_k}(x) = (x(t_1), \ldots, x(t_k)).$$

Since all $\lambda \in \Lambda$ fix 0 and T, then the projections π_0 and π_T are continuous. However, for $t \in (0,T)$ the projection π_t is continuous at x if and only if x is continuous at t. In addition, if we let $\mathcal{D}$ denote the Borel σ-algebra of $D_{\mathbb{R}^d}[0,T]$, then the projection $\pi_{t_1,\dots,t_k}$ is always measurable with respect to $\mathcal{D}$ and the Borel σ-algebra of $\mathbb{R}^k$.

We are particularly interested in the collections of finite-dimensional sets

$$\mathcal{F}_S = \left\{\pi^{-1}_{t_1,\dots,t_k}(H) : H \in \mathbb{R}^k; t_1 < \cdots < t_k \in S; k \geq 1\right\},$$

where S may be any subset of $[0,T]$. The reason is that these collections can be used to characterize probability measures on $\mathcal{D}$.

Theorem A.2.9. Let S be a dense subset of $[0,T]$ containing T. Then $\mathcal{F}_S$ generates the whole σ-algebra $\mathcal{D}$. In particular $\mathcal{F}_S$ is a separating class, in the sense that any two probability measures on $\mathcal{D}$, which agree on $\mathcal{F}_S$, are the same.

A.2.3 Weak convergence

For each $t \in [0,T]$ let D_t be the set of those càdlàg functions x such that π_t is discontinuous at x. Given a probability measure P on $\mathcal{D}$ we may now consider the set C_P of those $t \in [0,T]$ for which $P(D_t) = 0$. It is clear that $0, T \in C_P$, and we further have the following.

Proposition A.2.10. Let P be a probability measure on $\mathcal{D}$. Then $0, T \in C_P$ and the complement of C_P is countable.

Given two probability measures P and Q on $\mathcal{D}$, the last proposition tells us that $C_P \cap C_Q$ is a dense subset of $[0,T]$ that contains T. Therefore, by Theorem A.2.9, we know that $\mathcal{F}_{C_P \cap C_Q}$ is a separating class. Moreover, if $t_1 < \cdots < t_k$ lie in C_P and $P_n \Rightarrow P$ in $D_{\mathbb{R}^d}[0,T]$ as $n \to \infty$, then $P_n \pi^{-1}_{t_1,\dots,t_k} \Rightarrow P\pi^{-1}_{t_1,\dots,t_k}$ in $\mathbb{R}^k$ as $n \to \infty$, by Theorem A.1.5. Using Proposition A.1.11 we get the following converse.

Theorem A.2.11. If $(P_n)_{n\geq 1}$ is tight and $P_n \pi^{-1}_{t_1,\dots,t_k} \Rightarrow P\pi^{-1}_{t_1,\dots,t_k}$ in $\mathbb{R}^k$ as $n \to \infty$ for all $t_1 < \cdots < t_k$ lying in C_P, then $P_n \Rightarrow P$ in $D_{\mathbb{R}^d}[0,T]$ as $n \to \infty$.

Moreover, Theorem A.2.8 gives the following criteria for proving tightness.

Theorem A.2.12. A sequence $(P_n)_{n\geq 1}$ of probability measures on $\mathcal{D}$ is tight if and only if the following conditions hold.

1. For each $\eta > 0$ there exists $M > 0$ such that
$$\sup_{n\geq 1} P_n\left(\{x \in D_{\mathbb{R}^d}[0,T] : ||x|| \geq M\}\right) \leq \eta.$$
2. For each $\varepsilon, \eta > 0$ there exist $\delta \in (0,T)$ and $n_0 \geq 1$ such that
$$\sup_{n\geq n_0} P_n\left(\{x \in D_{\mathbb{R}^d}[0,T] : w'_x(\delta) \geq \varepsilon\}\right) \leq \eta.$$

A.3 The space $D_{\mathbb{R}^d}[0,\infty)$

We now define the space $D_{\mathbb{R}^d}[0,\infty)$ and establish the relation between weak convergence in the spaces $D_{\mathbb{R}^d}[0,T]$ and weak convergence in $D_{\mathbb{R}^d}[0,\infty)$.

Definition A.3.1. The space $D_{\mathbb{R}^d}[0,\infty)$ is the space of càdlàg functions on $[0,+\infty)$.

Given an interval $[0,t] \subset [0,+\infty)$ we let d_0^t denote the metric that we introduced in Definition A.2.6. All functions $x, y \in D_{\mathbb{R}^d}[0,\infty)$ may be restricted to a càdlàg function on $[0,t]$, and thus it makes sense to write $d_0^t(x,y)$ to denote the distance between their restrictions to $[0,t]$. In the sequel x and x_n always lie in $D_{\mathbb{R}^d}[0,\infty)$.

Proposition A.3.2. Suppose that $d_0^t(x_n,x) \to 0$ as $n \to \infty$. If $s \in (0,t)$ and x is continuous at s, then $d_0^s(x_n,x) \to 0$ as $n \to \infty$.

For each integer $m \geq 1$ we define

$$g_m(t) = \begin{cases} 1 & \text{if } t \in [0, m-1], \\ m-t & \text{if } t \in (m-1, m), \\ 0 & \text{if } t \in [m, +\infty). \end{cases}$$

For each $x \in D_{\mathbb{R}^d}[0,\infty)$ note that $x^m(t) = g_m(t)x(t)$ is continuous at m.

Definition A.3.3. For all $x, y \in D_{\mathbb{R}^d}[0,\infty)$ we define

$$d_0^\infty(x,y) = \sum_{m=1}^{\infty} \frac{1}{2^m} \min(1, d_0^m(x^m, y^m)).$$

It is easy to see that d_0^∞ is a metric. Moreover, d_0^∞ has the natural property that $d_0^\infty(x_n,x) \to 0$ implies, by Proposition A.3.2, that $d_0^t(x_n,x) \to 0$ as $n \to \infty$ whenever x is continuous at t; the converse is also true.

Theorem A.3.4. There is convergence $d_0^\infty(x_n,x) \to 0$ as $n \to \infty$ if and only if $d_0^t(x_n,x) \to 0$ as $n \to \infty$ for each continuity point t of x.

For each $x \in D_{\mathbb{R}^d}[0,\infty)$ let $\psi_m x$ be the restriction of x^m to $[0,m]$; this defines a continuous map $\psi_m : D_{\mathbb{R}^d}[0,\infty) \longrightarrow D_{\mathbb{R}^d}[0,m]$. Consider now the product space

$$\Pi = \prod_{m=1}^{\infty} D_{\mathbb{R}^d}[0,m],$$

whose elements will be denoted $\alpha = (\alpha_m)_{m \geq 1}$. The metric

$$\rho(\alpha,\beta) = \sum_{m=1}^{\infty} \frac{1}{2^m} \min(1, d_0^m(\alpha_m, \beta_m))$$

generates the product topology in Π; thus this metric makes Π separable and complete. Also, the map $\psi : D_{\mathbb{R}^d}[0,\infty) \longrightarrow \Pi$, such that $(\psi x)_m = \psi_m x$, is an isometry.

Theorem A.3.5. The image of $D_{\mathbb{R}^d}[0,\infty)$ under ψ is closed in Π. In particular, $(D_{\mathbb{R}^d}[0,\infty), d_0^\infty)$ is separable and complete.

It is possible to characterize the compact sets and the finite-dimensional sets of $D_{\mathbb{R}^d}[0,\infty)$ as it was done for $D_{\mathbb{R}^d}[0,T]$, and afterwards one may obtain criteria for proving the weak convergence of probability measures in $D_{\mathbb{R}^d}[0,\infty)$. We will however just state a result establishing the relation between the weak convergence in $D_{\mathbb{R}^d}[0,\infty)$ and the weak convergence in the spaces $D_{\mathbb{R}^d}[0,T]$.

Let $\mathcal{D}_\infty$ denote the Borel σ-algebra of $D_{\mathbb{R}^d}[0,\infty)$. Given a probability measure P on $\mathcal{D}_\infty$ we let D_t be the set of those $x \in D_{\mathbb{R}^d}[0,\infty)$ that are discontinuous at t, and we define C_P to be the set of those $t \geq 0$ such that $P(D_t) = 0$; this is the same definition that we gave in Subsection A.2.3. In addition, we define for each $t \geq 0$ the map $r_t : D_{\mathbb{R}^d}[0,\infty) \longrightarrow D_{\mathbb{R}^d}[0,t]$ that restricts each $x \in D_{\mathbb{R}^d}[0,\infty)$ to the interval $[0,t]$. It is possible to prove that these maps are Borel measurable; we moreover have the following theorem.

Theorem A.3.6. Consider probability measures P and P_n on $\mathcal{D}_\infty$. The weak convergence $P_n \Rightarrow P$ in $D_{\mathbb{R}^d}[0,\infty)$ as $n \to \infty$ occurs if and only if $P_n r_t^{-1} \Rightarrow P r_t^{-1}$ in $D_{\mathbb{R}^d}[0,t]$ as $n \to \infty$ for all $t \in C_P$.

Appendix B

Limit theorems for the Poisson process

In this appendix we state and prove some laws of large numbers, and a central limit theorem, for the Poisson process. We assume that the reader is familiar with the definitions of the Poisson and Wiener processes.

B.1 Laws of large numbers

Let $\mathcal{N}$ be a Poisson process with unitary intensity, defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The centered Poisson process Y is given by $Y(t) = \mathcal{N}(t) - t$.

Theorem B.1.1. Let α be a non-negative constant.

1. $\sup_{t\in[0,T]} \left| \frac{Y(nt)}{n^{1-\alpha}} \right| \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$ for all $\alpha \in [0, 1/4)$ and all $T \geq 0$.

2. $\sup_{t\in[0,T]} \left| \frac{Y(nt)}{n^{1-\alpha}} \right| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ for all $\alpha \in [0, 1/2)$ and all $T \geq 0$.

Proof. Since Y^4 is a submartingale, Doob's maximal inequality yields the following equation for each $\varepsilon > 0$.

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left| \frac{Y(nt)}{n^{1-\alpha}} \right| \geq \varepsilon \right) \leq \frac{\mathbb{E}\left[Y^4(nT)\right]}{\varepsilon^4 n^{4(1-\alpha)}} = \frac{nT + 3(nT)^2}{\varepsilon^4 n^{4(1-\alpha)}} = \frac{T}{\varepsilon^4 n^{3-4\alpha}} + \frac{3T^2}{\varepsilon^4 n^{2-4\alpha}}.$$

The right-hand side of the above equation converges to zero as $n \to \infty$ for all $\alpha \in [0, 1/2)$, and this observation proves the second claim. Furthermore, if we assume that $\alpha \in [0, 1/4)$, then

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \left| \frac{Y(nt)}{n^{1-\alpha}} \right| \geq \varepsilon \right) < \infty,$$

and the following observation completes the proof of the first claim.

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}\left\{\sup_{t\in[0,T]}\left|\frac{Y(nt)}{n^{1-\alpha}}\right|\geq\varepsilon\right\}\right)\leq\lim_{m\to\infty}\sum_{n=m}^{\infty}\mathbb{P}\left(\sup_{t\in[0,T]}\left|\frac{Y(nt)}{n^{1-\alpha}}\right|\geq\varepsilon\right)=0.$$

□

B.2 Central limit theorem

We prove the central limit theorem in a time inhomogeneous setting.

Theorem B.2.1. Consider an integrable and bounded map $f:[0,T]\longrightarrow[0,+\infty)$, a centered Poisson process Y with intensity one and a standard Wiener process W.

$$\frac{1}{\sqrt{n}}Y\left(\int_0^t nf(\tau)d\tau\right)\Rightarrow W\left(\int_0^t f(\tau)d\tau\right)\quad\text{in } D_{\mathbb{R}}[0,T]\quad\text{as } n\to\infty.$$

Proof. Define the processes

$$U_n(t)=\frac{1}{\sqrt{n}}Y\left(\int_0^t nf(\tau)d\tau\right)\quad\text{and}\quad U(t)=W\left(\int_0^t f(\tau)d\tau\right).$$

Also, using the notation of Subsection A.1.1, let $P_n=\mathbb{P}U_n^{-1}$ and $P=\mathbb{P}U^{-1}$. It is easy to see, using the continuous mapping theorem, that the convergence of

$$(U_n(t_1),U_n(t_2)-U_n(t_1),\ldots,U_n(t_k)-U_n(t_{k-1})),\tag{B.1}$$

in distribution as $n\to\infty$, to the random vector

$$(U(t_1),U(t_2)-U(t_1),\ldots,U(t_k)-U(t_{k-1}))\tag{B.2}$$

implies the weak convergence of the finite-dimensional distributions; namely, the limit in distribution $P_n\pi^{-1}_{t_1,\ldots,t_k}\Rightarrow P\pi^{-1}_{t_1,\ldots,t_k}$ in $\mathbb{R}^k$ as $n\to\infty$.

Note that the increments $U_n(t_i)-U_n(t_{i-1})$ are independent, as well as the increments $U(t_i)-U(t_{i-1})$. Moreover, by the central limit theorem for random variables

$$U_n(t_i)-U_n(t_{i-1})\sim\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Y_i\left(\int_{t_{i-1}}^{t_i}f(\tau)d\tau\right)\Rightarrow N\left(0,\int_{t_{i-1}}^{t_i}f(\tau)d\tau\right)\sim U(t_i)-U(t_{i-1})$$

as $n\to+\infty$, where $\{Y_i\}_{i\geq1}$ is an independent family of centered Poisson processes with unitary intensity. These two observations imply that the expression (B.1) converges to (B.2) for all $0\leq t_1<\cdots<t_k\leq T$. Hence, the finite-dimensional distributions of U_n converge to those of U.

Consequently, by Theorem A.2.11, it is now enough to show that the sequence $(U_n)_{n\geq1}$ is tight. To do this, by Theorem A.2.12, it suffices to prove that:

1. For each $\eta>0$ there exists $M>0$ such that

$$\mathbb{P}\left(\sup_{t\in[0,T]}|U_n(t)|\geq M\right)\leq\eta\quad\forall\ n\geq1.$$

2. For each $\varepsilon, \eta > 0$ there exist $\delta \in (0, T)$ and $n_0 \geq 1$ such that

$$\mathbb{P}\left(w'_{U_n}(\delta) \geq \varepsilon\right) \leq \eta \quad \forall\, n \geq n_0.$$

In order to check that the above conditions hold, define

$$J(t) = \int_0^t f(\tau) d\tau$$

and let $S = J(T)$. Since J is a non-decreasing continuous function with $J(0) = 0$, then the image of the interval $[0, T]$ under J is exactly $[0, S]$.

For the first condition, suppose that we are given some $\eta > 0$ and choose $M > 0$ such that $S + 3S^2 \leq M^4\eta$. Then, after applying Doob's maximal inequality to the submartingale Y^4, we see that

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,T]} |U_n(t)| \geq M\right) &\leq \mathbb{P}\left(\sup_{s\in[0,S]} |Y(ns)| \geq \sqrt{n}M\right) \\ &\leq \frac{nS + 3(nS)^2}{n^2M^4} \leq \frac{S + 3S^2}{M^4} \leq \eta \quad \forall\, n \geq 1.\end{aligned}$$

In order to prove that the second condition also holds, we are going to fix some $\varepsilon, \delta > 0$ and $n \geq 1$, and then we will provide a bound for

$$\mathbb{P}\left(w'_{U_n}(\delta) \geq \varepsilon\right).$$

To this end, consider the set Ψ_δ of all partitions $\pi = \{0 = t_0 < \cdots < t_{m_\pi} = T\}$, such that $\delta < t_{i+1} - t_i \leq 2\delta$ for all $i = 0, \ldots, m_\pi - 1$. Note that we may write

$$w'_{U_n}(\delta) = \inf_{\pi\in\Psi_\delta} \max_{0\leq i<m_\pi} w_{U_n}[t_i, t_{i+1}).$$

This expression is difficult to handle because it depends on all the partitions in Ψ_δ, we will thus provide a bound for this expression that we may handle more easily. To this purpose, choose a partition $0 = s_0 < \cdots < s_m = T$ such that $\delta \leq s_{i+1} - s_i \leq 2\delta$ for all $i = 0, \ldots, m - 1$. Also, assume that $\pi = \{0 = t_0 < \cdots < t_{m_\pi} = T\} \in \Psi_\delta$ and take two constants $0 \leq \alpha \leq \beta < T$, such that $t_i \leq \alpha \leq \beta < t_{i+1}$ for some $i \in \{0, \ldots, m_\pi - 1\}$. Since $\beta - \alpha < 2\delta$, we have three possible scenarios:

1. $s_k \leq \alpha \leq \beta \leq s_{k+1}$ for some $k = 0, \ldots, m - 1$.

2. $s_k \leq \alpha < s_{k+1} < \beta \leq s_{k+2}$ for some $k = 0, \ldots, m - 2$.

3. $s_k \leq \alpha < s_{k+1} < s_{k+2} < \beta \leq s_{k+3}$ for some $k = 0, \ldots, m - 3$.

We will derive the following inequality assuming that we are in the third case, but the same may be done in the other two cases.

$$\begin{aligned}|U_n(\beta) - U_n(\alpha)| &\leq |U_n(\beta) - U_n(s_{k+2})| + |U_n(s_{k+2}) - U_n(s_{k+1})| + |U_n(s_{k+1}) - U_n(s_k)| \\ &\quad + |U_n(s_k) - U_n(\alpha)| \leq 4 \max_{0\leq k<m} \sup_{\theta\in[0,2\delta]} |U_n(s_k + \theta) - U_n(s_k)|.\end{aligned}$$

In particular, this implies that

$$w_{U_n}[t_i, t_{i+1}) \leq 4 \max_{0 \leq k < m} \sup_{\theta \in [0,2\delta]} |U_n(s_k + \theta) - U_n(s_k)|.$$

The above is true for all $i = 0, \ldots, m_\pi - 1$ and for all $\pi \in \Psi_\delta$, and thus we have

$$\begin{aligned}
\mathbb{P}\left(w'_{U_n}(\delta) \geq \varepsilon\right) &\leq \mathbb{P}\left(\max_{0 \leq k < m} \sup_{\theta \in [0,2\delta]} |U_n(s_k + \theta) - U_n(s_k)| \geq \frac{\varepsilon}{4}\right) \\
&\leq \sum_{k=0}^{m-1} \mathbb{P}\left(\sup_{\theta \in [0,2\delta]} |U_n(s_k + \theta) - U_n(s_k)| \geq \frac{\varepsilon}{4}\right).
\end{aligned}$$

Let $M_f = \sup\{|f(t)| : t \in [0,T]\}$ and note that $|J(t) - J(s)| \leq M_f|t - s|$ for all $s, t \in [0,T]$. In particular, $|J(s_k + \theta) - J(s_k)| \leq M_f\theta$ and thus

$$\begin{aligned}
\mathbb{P}\left(w'_{U_n}(\delta) \geq \varepsilon\right) &\leq \sum_{k=0}^{m-1} \mathbb{P}\left(\sup_{\nu \in [0,2M_f\delta]} \frac{|Y\left(nJ(s_k) + n\nu\right) - Y(nJ(s_k))|}{\sqrt{n}} \geq \frac{\varepsilon}{4}\right) \\
&= m\mathbb{P}\left(\sup_{\nu \in [0,2M_f\delta]} \frac{|Y(n\nu)|}{\sqrt{n}} \geq \frac{\varepsilon}{4}\right),
\end{aligned}$$

Furthermore, since $m \leq T/\delta$, we see that

$$\begin{aligned}
m\mathbb{P}\left(\sup_{\nu \in [0,2M_f\delta]} \frac{|Y(n\nu)|}{\sqrt{n}} \geq \frac{\varepsilon}{4}\right) &\leq \frac{T}{\delta}\left(\frac{4}{\varepsilon}\right)^4 \frac{2M_f\delta n + 3(2M_f\delta n)^2}{n^2} \\
&= T\left(\frac{\varepsilon}{4}\right)^4 \left(\frac{2M_f}{n} + 12M_f^2\delta\right).
\end{aligned}$$

It is clear that if we are given some $\varepsilon, \eta > 0$, then we may choose $\delta > 0$ and $n_0 \geq 1$ such that the right-hand side is smaller than η for all $n \geq n_0$. □

Appendix C

Markov processes and infinitesimal generators

In this appendix we present, in a rather concise manner, several definitions and results regarding Markov processes and their characterization by means of infinitesimal generators. Section C.1 concerns semigroups of operators and their infinitesimal generators, and it is based on [8, Chapter 1]. Markov processes are introduced in Section C.2 which is entirely based on [8, Chapter 4]. Afterwards, we discuss Feller semigroups of operators, and the corresponding processes, in Section C.3, which is based on [16, Chapter 17] and [31, Chapter 7.1]. Feller diffusions are defined in Section C.4, which is also based on [16, Chapter 17].

C.1 Operator semigroups

Consider a Banach space $(M, ||\cdot||)$ and denote by $B(M)$ the set of all bounded linear operators $T : M \longrightarrow M$; we will use the notation $||\cdot||$ to denote the operator norm as well.

Definition C.1.1. An operator semigroup is a family $\{T_t\}_{t\geq 0} \subset B(M)$ such that

1. T_0 is the identity operator.

2. $T_s T_t = T_{s+t}$ for all $s, t \geq 0$.

An operator semigroup is strongly continuous if

$$\lim_{t\to 0} ||T_t f - f|| = 0 \quad \forall\ f \in M.$$

Besides, an operator semigroup is contractive if $||T_t|| \leq 1$ for all $t \geq 0$.

The basic example of a strongly continuous semigroup of operators is the expo-

nential of a bounded operator $B \in B(M)$, which is given by

$$e^{tB} = \sum_{n=0}^{\infty} \frac{t^n}{n!} B^n \quad \forall\ t \geq 0.$$

Furthermore, the bound

$$\left|\left|e^{tB}\right|\right| \leq \sum_{n=0}^{\infty} \frac{t^n}{n!} ||B||^n = e^{t||B||}$$

implies that $T_t = e^{-t||B||} e^{tB}$ is also a contraction semigroup. A similar inequality holds for strongly continuous semigroups in general.

Proposition C.1.2. Let $\{T_t\}_{t\geq 0}$ be a strongly continuous semigroup of operators on M. There exist constants $K, \alpha > 0$ such that

$$||T_t|| \leq K e^{\alpha t} \quad \forall\ t \geq 0.$$

Using the last proposition it is possible to prove following.

Proposition C.1.3. Let $\{T_t\}_{t\geq 0}$ be a strongly continuous semigroup of operators on M. For each $f \in M$ the map $[0, +\infty) \to M$ such that $t \mapsto T_t f$ is continuous.

A linear operator A on M is a linear mapping whose domain $\mathcal{D}(A)$ is a subspace of M; its range is denoted $\mathcal{R}(A)$. The graph of A is defined to be the set

$$\mathcal{G}(A) = \{(f, Af) : f \in \mathcal{D}(A)\} \subset M \times M.$$

The linear space $M \times M$, with componentwise addition and multiplication, is a Banach space if we endow it with the norm $|||(f, g)||| = ||f|| + ||g||$. A linear operator A is said to be closed if its graph is a closed subspace of $M \times M$.

Definition C.1.4. The infinitesimal generator of a semigroup of operators $\{T_t\}_{t\geq 0}$ on M is the linear operator A defined by the limit

$$Af = \lim_{t\to 0} \frac{1}{t}(T_t f - f),$$

whose domain is the subspace of all $f \in M$ such that the limit exists.

Before we can state some properties of infinitesimal generators, we need to discuss the calculus of functions taking values in Banach spaces. To this purpose, suppose that $I = [a, b]$ is a bounded and closed interval, and consider a set of points of the form $\pi = \{t_0 \leq s_1 \leq t_1 \leq \cdots \leq t_{n-1} \leq s_n \leq t_n\}$ such that the subset $\{a = t_0 < t_1 < \cdots < t_n = b\}$ is a partition of I; we define the norm of such a set by

$$||\pi|| = \max_{1\leq i\leq n} t_i - t_{i-1}$$

A function $u : I \longrightarrow M$ is Riemann integrable on I if the limit

$$\int_a^b u(t)dt = \lim_{||\pi||\to 0} \sum_{i=1}^{n} u(s_i)(t_i - t_{i-1})$$

exists. In the case of an unbounded closed interval, for instance $I = [a, +\infty)$, we say that a function $u : I \longrightarrow M$ is Riemann integrable on I if it is Riemann integrable on $[a, b]$ for all $b \geq a$ and the following limit exists.

$$\int_a^{+\infty} u(t)dt = \lim_{b \to +\infty} \int_a^b u(t)dt$$

We are going to let $C_M(I)$ denote the set of all continuous functions $u : I \longrightarrow M$, and we will let $C_M^1(I)$ be the set of all continuously differentiable functions; the derivative of u at $t \in I$ is defined as the limit

$$\frac{du}{dt}(t) = \lim_{h \to 0} \frac{1}{h}[u(t+h) - u(t)],$$

whenever the limit exists.

Lemma C.1.5. Consider a closed interval $I \subset \mathbb{R}$.

1. If $u \in C_M(I)$ and $||u||$ has a finite integral, then u is integrable over I, and
$$\left|\left|\int_I u(t)dt\right|\right| \leq \int_I ||u(t)||\, dt.$$
In particular, if I is bounded, then u is integrable over I.

2. Let A be a closed linear operator on M, and suppose that $u \in C_M(I)$. Furthermore, assume that $u(t) \in \mathcal{D}(A)$ for all $t \geq 0$, $Au \in C_M(I)$ and both u and Au are integrable over I. Then the integral of u over I belongs to $\mathcal{D}(A)$ and
$$A \int_I u(t)dt = \int_I Au(t)dt.$$

3. Assume that $I = [a, b]$ and $u \in C_M^1(I)$, then
$$\int_a^b \frac{du}{dt}(t)dt = u(b) - u(a).$$

We may now state the following properties.

Proposition C.1.6. Let $\{T_t\}_{t \geq 0}$ be a strongly continuous semigroup of operators on M with infinitesimal generator A.

1. If $f \in M$ and $t \geq 0$, then the integral of $T_t f$ over $[0, t]$ belongs to $\mathcal{D}(A)$, and
$$T_t f - f = A \int_0^t T_s f ds.$$

2. If $f \in \mathcal{D}(A)$ and $t \geq 0$, then $T_t f \in \mathcal{D}(A)$, and
$$\frac{dT_t f}{dt} = AT_t f = T_t Af.$$

3. If $f \in \mathcal{D}(A)$ and $t \geq 0$, then
$$T_t f - f = \int_0^t AT_s f ds = \int_0^t T_s Af ds.$$

Note that the integrals exist because the map $t \mapsto T_t f$ is continuous for all $f \in B(M)$. Moreover, using the last proposition it is possible to prove the following.

Proposition C.1.7. If A is the infinitesimal generator of a strongly continuous semigroup of operators on M, then $\mathcal{D}(A)$ is dense in M and A is closed.

Given a closed linear operator A on M, we say that $\lambda \in \mathbb{R}$ belongs to the resolvent set $\rho(A)$ if the map $\lambda - A$ is injective, the range $\mathcal{R}(\lambda - A) = M$ and the inverse $(\lambda - A)^{-1}$ is a bounded operator. In that case we say that $R_\lambda = (\lambda - A)^{-1}$ is the resolvent operator at λ.

In general, for any linear closed operator A, the fact that the maps $(\lambda - A)$ and $(\mu - A)$ commute for all $\lambda, \mu \in \mathbb{R}$ implies that their inverses commute as well, when they exist. Furthermore, we have the identity

$$R_\lambda R_\mu = \frac{1}{\lambda - \mu}(R_\mu - R_\lambda) = R_\mu R_\lambda \quad \forall\ \lambda, \mu \in \rho(A).$$

Also, if $\lambda \in \rho(A)$ and $|\lambda - \mu| < ||R_\lambda||^{-1}$, then

$$\sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^{n+1}$$

defines a bounded operator that is in fact $(\mu - A)^{-1}$. This implies that $\rho(A)$ is open.

In the special case of the infinitesimal generator A of a strongly continuous contraction semigroup of operators $\{T_t\}_{t \geq 1}$ on M, we see that for each $\lambda > 0$ the linear map

$$U_\lambda g = \int_0^{+\infty} e^{-\lambda t} T_t g dt$$

is a bounded operator on M. Indeed, the integrand belongs to $C_M([0, +\infty))$ and the exponential, together with the fact that $\{T_t\}_{t \geq 0}$ is a contraction semigroup, ensure that the integral converges. Moreover, the contraction property also implies, by the first item of Lemma C.1.5, that $||U_\lambda g|| \leq \lambda^{-1} ||g||$.

Proposition C.1.8. Let $\{T_t\}_{t \geq 0}$ be a strongly continuous contraction semigroup of operators on M with infinitesimal generator A. Then $(0, +\infty) \subset \rho(A)$, and moreover

$$R_\lambda g = U_\lambda g = \int_0^{+\infty} e^{-\lambda t} T_t g dt \quad \forall\ g \in M,\ \lambda > 0.$$

A linear operator A is said to be dissipative if $||\lambda f - Af|| \geq \lambda ||f||$ for each $f \in \mathcal{D}(A)$ and each $\lambda > 0$. We now state a version of the Hille-Yosida theorem.

Theorem C.1.9 (Hille-Yosida theorem)**.** A linear operator A on M is the generator of a strongly continuous contraction semigroup on M if and only if

1. $\mathcal{D}(A)$ is dense in M.
2. A is dissipative.
3. There exists $\lambda > 0$ such that the range $\mathcal{R}(\lambda - A) = M$.

Note that the necessity is a consequence of the results that we have already seen; the fact that infinitesimal generators are dissipative results from the inequality $||R_\lambda g|| \leq \lambda^{-1} ||g||$, which was proven above.

Another very important fact about infinitesimal generators is the following.

Theorem C.1.10. Let $\{T_t\}_{t\geq 0}$ and $\{S_t\}_{t\geq 0}$ be strongly continuous contraction semigroups of operators on M, with infinitesimal generators A and B, respectively. If $A = B$, then $T_t = S_t$ for all $t \geq 0$.

C.2 Markov processes

Consider a complete and separable metric space E, and let $\mathcal{B}(E)$ denote its Borel σ-algebra. We are going to let $B(E)$ be the space of bounded measurable functions $f : E \longrightarrow \mathbb{R}$, which is a Banach space if we endow it with the norm

$$||f|| = \sup_{x\in E} |f(x)|.$$

Fix some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given an stochastic process $\{X_t\}_{t\geq 0}$, defined over this space, and taking values in E, we are going to let $\mathcal{F}_t = \sigma(\{X_s : s \geq t\})$.

Definition C.2.1. An stochastic process X is a Markov process if

$$\mathbb{P}(X_{s+t} \in \Gamma | \mathcal{F}_s) = \mathbb{P}(X_{s+t} \in \Gamma | X_s) \quad \text{a.s.}$$

for all $s, t \geq 0$ and all $\Gamma \subset \mathcal{B}(E)$. The latter is called the Markov property and it means that, given the present, the future does not depend on the past.

An equivalent formulation of the Markov property is

$$\mathbb{E}[f(X_{s+t})|\mathcal{F}_s] = \mathbb{E}[f(X_{s+t})|X_s] \quad \text{a.s.}$$

for all $s, t \geq 0$ and all $f \in B(E)$.

Definition C.2.2. A function $P : [0, +\infty) \times E \times \mathcal{B}(E) \longrightarrow [0, 1]$ is a time homogeneous transition function if it poses the following properties.

1. $P(t, x, \cdot\,)$ is a probability measure for all $t \in [0, +\infty)$ and all $x \in E$.

2. $P(0, x, \cdot\,) = \delta_x$, the unit mass at x for all $x \in E$.

3. $P(\cdot\,, \cdot\,, \Gamma)$ is Borel measurable for all $\Gamma \in \mathcal{B}(E)$.

4. For all $s, t \geq 0$, $x \in E$ and $\Gamma \in \mathcal{B}(E)$

$$P(s + t, x, \Gamma) = \int_E P(t, y, \Gamma) P(s, x, dy),$$

which is called the Chapman-Kolmogorov property.

Furthermore, we say that P is the transition function of the time homogeneous Markov process X if for all $s, t \geq 0$ and $\Gamma \in \mathcal{B}(E)$ we have

$$\mathbb{P}(X_{s+t} \in \Gamma | \mathcal{F}_s) = P(t, X_s, \Gamma) \quad \text{a.s.}$$

Intuitively, the meaning of $P(t, x, \Gamma)$ is the probability that $X_t \in \Gamma$ given that the initial state of X was $X_0 = x$.

Again, we observe that the last equation is equivalent to

$$\mathbb{E}[f(X_{s+t})|\mathcal{F}_s] = \int_E f(x) P(t, X_s, dx) \quad \text{a.s.} \tag{C.1}$$

for all $s, t \geq 0$ and all $f \in B(E)$.

We will often write $P_t(x, \Gamma)$ instead of $P(t, x, \Gamma)$.

Definition C.2.3. A probability measure ν on $\mathcal{B}(E)$ is said to be the initial distribution of the Markov process X if

$$\mathbb{P}(X_0 \in \Gamma) = \nu(\Gamma) \quad \forall\, \Gamma \in \mathcal{B}(E).$$

An important property is that a transition function and an initial distribution for a Markov process X determine its finite-dimensional distributions.

Proposition C.2.4. Let P and ν be, respectively, a transition function and an initial distribution for the Markov process X, then

$$\mathbb{P}(X_{t_1} \in \Gamma_1, \ldots, X_{t_n} \in \Gamma_n) = \int_E \nu(dx_0) \int_{\Gamma_1} P_{t_1}(x_0, dx_1) \ldots \int_{\Gamma_n} P_{t_n - t_{n-1}}(x_{n-1}, dx_n)$$

for all $0 \leq t_1 < \cdots < t_n$ and $\Gamma_1, \ldots, \Gamma_n \in \mathcal{B}(E)$.

In particular, this allows to prove the following.

Theorem C.2.5. Let P and ν be, respectively, a transition function and a probability measure on $\mathcal{B}(E)$. There exists a Markov process X whose transition function and initial distribution are P and ν, respectively.

For each $x \in E$ we are going to let P_ν denote the probability measure on the product σ-algebra $\otimes_{[0,+\infty)} \mathcal{B}(E)$, associated to the Markov process X in the statement of the last theorem; in the special case when the initial distribution is $\nu = \delta_x$ we use the notation P_x instead.

C.2.1 Operator semigroups and Markov processes

In general it is not possible to define transition functions explicitly. However, we may instead exploit the fact that

$$T_t f(x) = \int_E f(y) P_t(x, dy)$$

defines a contraction semigroup of operators on $B(E)$, whenever P is a transition function. Indeed, by the Chapman-Kolmogorov property we have

$$T_s T_t f(x) = \int_E \int_E f(z) P_t(y, dz) P_s(x, dy) = \int_E f(z) P_{s+t}(x, dz) = T_{s+t} f(x)$$

for all $s, t \geq 0$ and $f \in B(E)$. Also note that $T_0 = \mathrm{Id}$ because $P_0(x, \cdot\,) = \delta_x$, and that, since $P_t(x, \cdot\,)$ is a probability measure, then $||T_t|| \leq 1$.

Definition C.2.6. An operator semigroup $\{T_t\}_{t \geq 0}$ defined on some closed subspace $M \subset B(E)$ corresponds to a Markov process X if

$$\mathbb{E}[f(X_{s+t}) | \mathcal{F}_s] = T_t f(X_s) \quad \text{a.s.} \tag{C.2}$$

for all $s, t \geq 0$ and $f \in M$.

The above is intended to be the analog of equation (C.1), in the definition of the transition function of a Markov process; in particular, note that when $\{T_t\}_{t \geq 0}$ comes from a transition function then equation (C.2) holds.

We say that $M \subset B(E)$ is separating if for all $x, y \in E$ there exists $f \in M$ such that $f(x) \neq f(y)$. The following is an important fact.

Proposition C.2.7. Let X be a Markov process with initial distribution ν, corresponding to a operator semigroup $\{T_t\}_{t \geq 0}$, defined on M. If M is separating, then ν and $\{T_t\}_{t \geq 0}$ determine the finite-dimensional distributions of X.

Suppose that the initial distribution of a Markov process X is given, and that this process corresponds to some strongly continuous contraction semigroup of operators $\{T_t\}_{t \geq 0}$, defined on a closed and separating subspace $M \subset B(E)$. Then its infinitesimal generator A determines the finite-dimensional distributions of X by Theorem C.1.10. In the general case, when $\{T_t\}_{t \geq 0}$ is a generic semigroup of operators, it is necessary to consider its full generator, or a sufficiently large subset of it; however we will not discuss this problem here.

C.3 Feller processes

Assume that the space E in the latter section is a complete, separable and locally compact metric space, and let $C_0(E)$ denote the closed and separating subspace of $B(E)$ consisting of all continuous functions that vanish at infinity. Specifically, $f \in C_0(E)$ if f is continuous and for all $\varepsilon > 0$ there exists a compact set $K \subset E$ such that $|f(x)| < \varepsilon$ for all $x \notin K$.

Definition C.3.1. A contraction semigroup of operators $\{T_t\}_{t \geq 0}$ on $C_0(E)$ is said to be a Feller semigroup if

1. $\{T_t\}_{t \geq 0}$ is positive, meaning that $T_t f \geq 0$ for all $f \geq 0$.
2. For each $x \in E$ and each $f \in C_0(E)$ we have

$$\lim_{t \to 0} T_t f(x) = f(x).$$

Here we are considering a pointwise limit, but as it is stated below Feller semigroups are actually strongly continuous as well.

A Markov process with a transition function that defines a Feller semigroup of operators is called a Feller process.

Theorem C.3.2. Feller semigroups of operators are strongly continuous. In particular the infinitesimal generator of a Feller semigroup of operators completely determines the semigroup.

A linear operator A on a Banach space M is said to be closable if there exists a linear operator $\overline{A}$, called closure of A, such that $\overline{\mathcal{G}(A)} = \mathcal{G}(\overline{A})$. We may now state the following analog of the Hille-Yosida theorem for Feller semigroups of operators.

Theorem C.3.3. Let A be a linear operator on $C_0(E)$. Then A is closable and its closure is the infinitesimal generator of a Feller semigroup if and only if

1. $\mathcal{D}(A)$ is dense in $C_0(E)$.
2. If $\sup\{f^+(y) : y \in E\} \leq f(x)$ for some $f \in \mathcal{D}(A)$ and $x \in E$, then $Af(x) \leq 0$.
3. There exists $\lambda > 0$ such that $\mathcal{R}(\lambda - A)$ is dense in $C_0(E)$.

The second condition is known as the positive maximum principle.

C.3.1 Feller processes

Consider a Feller semigroup of operators $\{T_t\}_{t\geq 0}$. For each $t \geq 0$ and $x \in E$, the map $f \mapsto T_t f(x)$ is a linear functional on $C_0(E)$. Then, by the Riesz representation theorem, there exists a Borel measure $P_t(x, \cdot\,)$ such that

$$T_t f(x) = \int_E f(y) P_t(x, dy) \quad \forall\ f \in C_0(E).$$

Since T_t is a positive contraction, then $P_t(x, \cdot\,)$ is a non-negative and finite measure. However, $P_t(x, \cdot\,)$ may not be a probability measure, because its total variation might be smaller than one. In order to ensure that $P_t(x, \cdot\,)$ is a probability measure, we need T_t to be conservative, namely we must require that

$$\sup_{||f||\leq 1} T_t f(x) = 1 \quad \forall\ x \in E.$$

This is equivalent to requesting that $f \mapsto T_t f(x)$ has norm one; by Riesz's theorem the norm of this map equals the total variation of the corresponding measure.

To avoid restricting ourselves to conservative semigroups, we may consider the one-point compactification $E \cup \{\Delta\}$ of E, and the space $C(E \cup \{\Delta\})$ of continuous functions on $E \cup \{\Delta\}$. Each $f \in C_0(E)$ may be extended to a function in $C(E \cup \{\Delta\})$ by setting $f(\Delta) = 0$, and we may moreover extend each T_t defining

$$\hat{T}_t f = f(\Delta) + T_t(f - f(\Delta)) \quad \forall\ f \in C(E \cup \{\Delta\}).$$

This results in a positive contraction and strongly continuous semigroup of operators which is furthermore conservative, and it is possible to construct a transition function from this semigroup.

Theorem C.3.4. Let $\{T_t\}_{t\geq 0}$ be a Feller semigroup of operators. There exists a unique transition function P on $E \cup \{\Delta\}$ such that

$$T_t f(x) = \int_E f(y) P_t(x, dy)$$

for all $x \in E$ and $f \in C_0(E)$.

As a result, a Feller semigroup $\{T_t\}_{t\geq 0}$ and a probability ν on $\mathcal{B}(E \cup \{\Delta\})$ define a unique Markov process taking values in $E \cup \Delta$. Furthermore, we have the following result concerning the regularity of paths.

Theorem C.3.5. Let X be the process in $E \cup \Delta$ determined by the Feller semigroup $\{T_t\}_{t\geq 0}$ and the initial distribution ν. There exists a version $\tilde{X}$ of this process that has càdlàg paths and is such that $X_{s^-} = \Delta$ or $X_s = \Delta$ imply $\tilde{X}_t = \Delta$ for all $t \geq s$. Moreover, if $\{T_t\}_{t\geq 0}$ is conservative and ν can be restricted to a probability measure on E, then $\tilde{X}$ can be chosen to be a càdlàg process in E.

Suppose that X is a Feller process in $E \cup \Delta$ with the properties that we stated in the last theorem. Then we may define the explosion time of X to be

$$\zeta = \inf \{t \geq 0 : X_{t^-} = \Delta \text{ or} X_t = \Delta\},$$

and this results in $X_t = \Delta$ for all $t \geq \zeta$.

We conclude this section with the Dynkin formula.

Theorem C.3.6. Consider a right-continuous Feller process X with infinitesimal generator A and initial distribution ν. Then the process

$$f(X_t) - f(X_0) - \int_0^t Af(X_s)ds$$

is a martingale with respect to its natural filtration and P_ν, for all $f \in \mathcal{D}(A)$; furthermore in the especial case $\nu = \delta_x$, where $x \in E$, we have the Dynkin formula

$$\mathbb{E}_x[f(X_t)] = f(x) + \mathbb{E}_x\left[\int_0^t Af(X_s)ds\right].$$

We also have the following reverse formula.

Theorem C.3.7. Let X be a Feller process with infinitesimal generator A. Suppose that $f, g \in C_0(E)$ are such that

$$f(X_t) - f(x) - \int_0^t g(X_s)ds$$

is a martingale with respect to the natural filtration of X and P_x, for each $x \in E$. Then $f \in \mathcal{D}(A)$ and $Af = g$.

C.4 Feller diffusions

In this section we let $E = \mathbb{R}^d$ and we consider the space $C_c^\infty(\mathbb{R}^d)$ of all infinitely differentiable functions with compact support. A linear operator in $C_c^\infty(\mathbb{R}^d)$ is said to be local if $Af(x) = 0$ whenever f vanishes in some neighborhood of x.

Theorem C.4.1. Let A be the infinitesimal generator of a Feller process X and assume that $C_c^\infty(\mathbb{R}^d) \subset \mathcal{D}(A)$. Then X is P_ν-almost surely continuous on $[0, \zeta)$, for each initial distribution ν, if and only if A is local; here ζ is the explosion time of X. Moreover, in the latter case there exist continuous functions $a_i, \sigma_{i,j}^2, c : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $\sigma^2 = (\sigma_{i,j}^2)$ is a symmetric positive semi-definite matrix and

$$Af = \sum_{i=1}^{d} a_i \frac{\partial f}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{d} \sigma_{i,j}^2 \frac{\partial^2 f}{\partial x_i \partial x_j} - cf \quad \forall\ f \in C_c^\infty(\mathbb{R}^d).$$

Consider now the second order differential operator L on $C_c^\infty(\mathbb{R}^d)$ such that

$$Lf = \sum_{i=1}^{d} a_i \frac{\partial f}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{d} \sigma_{i,j}^2 \frac{\partial^2 f}{\partial x_i \partial x_j} \quad \forall\ f \in C_c^\infty(\mathbb{R}^d), \tag{C.3}$$

where a and σ^2 are as above; note that we have taken $c = 0$. The functions a and σ^2 are called, respectively, the drift and the diffusion coefficients; the dispersion coefficient is defined as the square root of the positive semi-definite matrix σ^2.

Definition C.4.2. A Feller diffusion is a Feller process with continuous paths and such that the restriction of its infinitesimal generator to $C_c^\infty(\mathbb{R}^d)$ is a second order differential operator as the one above.

The following appendix provides a means of constructing diffusions, using stochastic differential equations.

Appendix D

Itô calculus and stochastic differential equations

The first part of this appendix concerns the definition of Itô integrals and their most relevant properties, including the Itô formula; this is covered in Section D.1 which is based on [4, Chapter 2] and [30, Chapter 4]. The second part of this appendix is devoted to stochastic differential equations and their invariant measures. In Section D.2 we give criteria for the existence and uniqueness of strong solutions to these equations, and we observe that these solutions are Feller diffusions; this section is based on [17, Chapter 5.2] and [33, Chapter 5.2]. Invariant measures are defined in Section D.3 where we further give criteria for their existence, uniqueness and ergodicity; this is based on [1, 7, 24, 25]

D.1 Itô calculus

Consider a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, endowed with a right-continuous and complete filtration $\{\mathcal{F}_t\}_{t\geq 0}$. This means that $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$ for all $t \geq 0$ and $\mathcal{F}_0$ contains the null sets. Suppose in addition that W is a standard $\mathcal{F}_t$-measurable Wiener process, such that for all $t \geq s$ the increment $W_t - W_s$ is independent of the past $\mathcal{F}_s$. For instance, this is the case when $\{\mathcal{F}_t\}_{t\geq 0}$ is the augmentation of the natural filtration $\mathcal{G}_t = \sigma(\{W_s : s \in [0,t]\})$.

We would like to give some meaning to the stochastic integral

$$\int_0^t f(s) dW_s,$$

where f is a well-behaved stochastic process. Since the Wiener process has infinite variation on bounded intervals, it is not possible to define this integral pathwise by means of classical integration theory.

Let us fix the domain of integration to be the interval $[0, T]$. The class of integrands that we are going to consider is the set $\mathcal{H}_2[0, T]$ of all progressively measurable

processes $f : [0,T] \times \Omega \longrightarrow \mathbb{R}$ such that

$$\int_0^T \mathbb{E}[f^2(\tau)]d\tau < \infty.$$

The progressive measurability of f means that the restriction $f|_{[0,t]\times\Omega}$ is measurable with respect to the product σ-algebra $\mathcal{B}_t \otimes \mathcal{F}_t$ for all $t \in [0,T]$; here $\mathcal{B}_t$ denotes the Borel σ-algebra of the interval $[0,t]$.

We remark that $\mathcal{H}_2[0,T]$ is a closed subspace of $L^2([0,T]\times\Omega)$ and thus a Hilbert space. The stochastic integral will first be defined on a class of elementary processes which are dense in $\mathcal{H}_2[0,T]$ and this fact will then be used to define the stochastic integral in all $\mathcal{H}_2[0,T]$.

Definition D.1.1. A simple process is a stochastic process of the form

$$\bar{f}(t) = \sum_{k=0}^{m-1} f_k \mathbb{1}_{[t_k,t_{k+1})}(t) \quad t \in [0,T],$$

where $0 = t_0 < \cdots < t_m = T$, the random variables f_k are $\mathcal{F}_{t_k}$-measurable and $\mathbb{E}[f_k^2] < \infty$ for all $k = 0, \ldots, m-1$. Note that $\bar{f} \in \mathcal{H}_2[0,T]$.

The Itô integral of $\bar{f}$ is defined to be the following random variable.

$$\int_0^T \bar{f}(\tau)dW_\tau = \sum_{k=0}^{m-1} f_k(W_{t_{k+1}} - W_{t_k}).$$

It is easy to check that the Itô integral is linear within the class of simple processes. Moreover, using the independence of the increments of W with respect to the past, we see that this integral has mean zero and variance

$$\mathbb{E}\left[\left(\int_0^T \bar{f}(\tau)dW_\tau\right)^2\right] = \int_0^T \mathbb{E}\left[\bar{f}^2(\tau)\right]d\tau.$$

In other words, the norm of the Itô integral, as an element of $L^2(\Omega)$, is the same as the norm of the integrand as an element of $L^2([0,T]\times\Omega)$. This isometry is key for extending the definition of the Itô integral to other processes in $\mathcal{H}_2[0,T]$, but first we need the following result.

Proposition D.1.2. The set of simple processes is dense in $\mathcal{H}_2[0,T]$. Specifically, for each $f \in \mathcal{H}_2[0,T]$ there exists a sequence of simple processes $\bar{f}_n$ such that

$$\lim_{n\to\infty} \int_0^T \mathbb{E}\left[\left(f(\tau) - \bar{f}_n(\tau)\right)^2\right]d\tau = 0$$

The last proposition can be used to define the stochastic integral of a generic process $f \in \mathcal{H}_2[0,T]$. According to the proposition, we know that there exists a sequence of simple processes $\bar{f}_n$ that converge to f in $L^2([0,T]\times\Omega)$ and the isometry property implies that

$$\mathbb{E}\left[\left(\int_0^T \bar{f}_m(\tau)dW_\tau - \int_0^T \bar{f}_n(\tau)dW_\tau\right)^2\right] = \int_0^T \mathbb{E}\left[\left(\bar{f}_m(\tau) - \bar{f}_n(\tau)\right)^2\right]d\tau.$$

Hence, the sequence formed by the Itô integrals of the processes $\bar{f}_n$ is a Cauchy sequence in $L^2(\Omega)$. The Itô integral of f is defined to be the limit of this sequence:

$$\int_0^T \bar{f}_n(\tau)dW_\tau \xrightarrow{L^2} \int_0^T f(\tau)dW_\tau.$$

The limit is independent of the sequence of simple processes that we choose.

Proposition D.1.3. The Itô integral has the following properties.

1. The integral is linear with respect to the integrand.

2. $\mathbb{E}\left[\int_0^T f(\tau)dW_\tau\right] = 0 \quad \forall\ f \in \mathcal{H}_2[0,T].$

3. $\mathbb{E}\left[\left(\int_0^T f(\tau)dW_\tau\right)^2\right] = \int_0^T \mathbb{E}\left[f^2(\tau)\right]d\tau \quad \forall\ f \in \mathcal{H}_2[0,T].$

4. If $f_n \to f$ in $\mathcal{H}_2[0,T]$, with respect to the norm of $L^2([0,T]\times\Omega)$, then
$$\int_0^T f_n(\tau)dW_\tau \xrightarrow{L^2} \int_0^T f(\tau)dW_\tau.$$

5. If ξ is a bounded and $\mathcal{F}_s$-measurable random variable and $t > s$, then
$$\int_0^T \xi \mathbb{1}_{[s,t)}(\tau)f(\tau)dW_\tau = \xi\int_0^T \mathbb{1}_{[s,t)}(\tau)f(\tau)dW_\tau \quad \text{a.s.} \quad \forall\ f \in \mathcal{H}_2[0,T].$$

D.1.1 Stochastic integrals with variable upper limit

Let us define for each $t \in [0,T]$ the integral

$$\int_0^t f(\tau)dW_\tau = \int_0^T \mathbb{1}_{[0,t)}(\tau)f(\tau)dW_\tau.$$

We would like to view this integral as a random function of its upper limit. Recall that the Itô integral is defined as a limit in $L^2(\Omega)$ and it is thus uniquely determined up to a set of probability zero, that depends on the integrand. The problem is that the integrands in the above expression depend on the time parameter t, that ranges in the uncountable set $[0,T]$. Therefore, the stochastic integral with variable upper limit could a priori not be determined as a function of t in a set of positive probability.

For a simple process $\bar{f}$ we may write

$$\int_0^t \bar{f}(\tau)dW_\tau = \int_0^t \mathbb{1}_{[0,t)}(\tau)\bar{f}(\tau)dW_\tau = \sum_{k=0}^{m-1} f_k\left(W_{t\wedge t_{k+1}} - W_{t\wedge t_k}\right).$$

In this case the stochastic integral with variable upper limit is a well-defined stochastic process which is moreover a continuous martingale. The fact that simple processes are dense in $\mathcal{H}_2[0,T]$ allows to prove the following theorem.

Theorem D.1.4. Consider a process $f \in \mathcal{H}_2[0,T]$. There exists a martingale I with almost surely continuous paths such that

$$I_t = \int_0^t f(\tau) dW_\tau \quad \text{a.s.} \quad \forall\, t \in [0,T].$$

In the sequel, whenever we consider the integral

$$\int_0^t f(\tau) dW_\tau$$

as a function of its upper limit, we will assume that the corresponding stochastic process is given by its continuous martingale version I.

The fact that the Itô integral is a martingale implies that Doob's maximal inequality holds, and this allows to prove an important property of stochastic integrals. Namely, suppose that $f_n \to f$ in $\mathcal{H}_2[0,T]$, in other words

$$\lim_{n\to\infty} \int_0^T \mathbb{E}\left[\left(f(\tau) - f_n(\tau)\right)^2\right] d\tau = 0.$$

Then the following limit holds.

$$\sup_{t\in[0,T]} \left| \int_0^t f(\tau) dW_\tau - \int_0^t f_n(\tau) dW_\tau \right| \xrightarrow{L^2} 0 \quad \text{as } n \to \infty.$$

D.1.2 Extension of the class of integrands

The condition that the integrand must have a finite second moment is rather restrictive. Thus, we will extend the definition of the Itô integral to the class $\mathcal{L}_2[0,T]$ of progressively measurable processes $f : [0,T] \times \Omega \longrightarrow \mathbb{R}$ such that

$$\mathbb{P}\left(\int_0^T f^2(\tau) d\tau < \infty \right) = 1.$$

Simple processes within $\mathcal{L}_2[0,T]$ are defined as in Definition D.1.1, except that we do not require anymore that the second moments of the random variables f_k are finite. The Itô integral of a simple process is also defined as before, specifically

$$\int_0^T \bar{f}(\tau) dW_\tau = \sum_{k=0}^{m-1} f_k (W_{t_{k+1}} - W_{t_k}).$$

In order to extend the definition of the Itô integral, we once more approximate a generic integrand by simple processes.

Proposition D.1.5. For any $f \in \mathcal{L}_2[0,T]$ there exists a sequence $\bar{f}_n$ of simple processes in $\mathcal{L}_2[0,T]$ such that

$$\int_0^T \left(f(\tau) - \bar{f}_n(\tau)\right)^2 d\tau \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to \infty.$$

It is possible to prove that

$$\sup_{t\in[0,T]}\left|\int_0^t \bar{f}_m(\tau)dW_\tau - \int_0^t \bar{f}_n(\tau)d\tau\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as } m,n\to\infty.$$

In other words, the stochastic integrals of the simple processes form a Cauchy sequence and thus we may define the Itô integral of f as the limit of this sequence:

$$\sup_{t\in[0,T]}\left|\int_0^t f(\tau)dW_\tau - \int_0^t \bar{f}_n(\tau)dW_\tau\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as } n\to\infty.$$

We remark that this limit does not depend on the choice of the simple processes.

It is easy to check that the Itô integrals of simple processes are continuous with respect to the upper limit. Since convergence in probability implies almost sure convergence of a subsequence, the Itô integral of a generic process in $\mathcal{L}_2[0,T]$ is almost surely the uniform limit of continuous processes, and hence it is a continuous process itself. Nevertheless, the Itô integral of a process in $\mathcal{L}_2[0,T]$ may not be a martingale anymore.

An important property of these Itô integrals is the following. Suppose that

$$\int_0^T \left(f(\tau)-f_n(\tau)\right)^2 d\tau \xrightarrow{\mathbb{P}} 0 \quad \text{as } n\to\infty,$$

where f and f_n belong to $\mathcal{L}_2[0,T]$ for all $n\geq 1$. Then

$$\sup_{t\in[0,T]}\left|\int_0^t f(\tau)dW_\tau - \int_0^t f_n(\tau)dW_\tau\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as } n\to\infty.$$

D.1.3 The Itô formula

As it happens with Riemann integrals, the definition of the Itô integral is not very useful for handling these stochastic integrals. To compute Riemann integrals we usually resort to differentiation techniques: the Barrow rule, the chain rule or the integration by parts formula, but in the context of Itô calculus there is no differentiation theory. However, there exists an integral analog of the chain rule that turns out to be a very powerful tool, namely the Itô formula.

Let $a,b:[0,T]\times\Omega\longrightarrow\mathbb{R}$ be progressively measurable processes such that

$$\int_0^T |a(\tau)|d\tau<\infty \quad \text{a.s.} \quad \text{and} \quad \int_0^T b^2(\tau)d\tau<\infty \quad \text{a.s.}$$

Note that $b\in\mathcal{L}_2[0,T]$. Suppose in addition that X_0 is $\mathcal{F}_0$-measurable and let

$$X_t = X_0 + \int_0^t a(\tau)d\tau + \int_0^t b(\tau)dW_\tau \quad t\in[0,T].$$

The latter is denoted by means of the stochastic differential

$$dX_t = a(t)dt + b(t)dW_t.$$

The Itô formula tells us how smooth functions act on processes of this kind.

Theorem D.1.6 (Itô formula)**.** Suppose that X is as above and let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a twice continuously differentiable function, then

$$df(X_t) = \left[f'(X_t)a(t) + \frac{1}{2}b^2(t)f''(X_t) \right] dt + f'(X_t)b(t)dW_t.$$

Under the convention that $dt \cdot dW_t$, $dW_t \cdot dt$ and $(dt)^2$ are all equal to zero, and $(dW_t)^2 = dt$, we may express the above stochastic differential as follows.

$$df(X_t) = f'(X_t)dX_t + \frac{1}{2}f''(X_t)(dX_t)^2.$$

This mnemonic device has a mathematical foundation; for instance, $(dW_t)^2 = t$ comes from the fact that the quadratic variation of W_t is t.

Consider now a d-dimensional Wiener process W, with independent coordinates, that is adapted to $\{\mathcal{F}_t\}_{t\geq 0}$ with increments that are independent of the past. Also, suppose that X is a d-dimensional vector random process with coordinates

$$dX_t^i = a_i(t)dt + b_{i,1}(t)dW_t^1 + \cdots + b_{i,n}(t)dW_t^n,$$

where the coefficients a_i and $b_{i,j}$ are progressively measurable processes such that

$$\int_0^T |a_i(\tau)|d\tau < \infty \quad \text{a.s.} \quad \text{and} \quad \int_0^T b_{i,j}^2(\tau)d\tau < \infty \quad \text{a.s.}$$

for all $i, j = 1, \ldots, d$. If we let $a = (a_i)$ and $b = (b_{i,j})$, then we may return to the more compact notation

$$dX_t = a(t)dt + b(t)dW_t.$$

Theorem D.1.7. Suppose that X is as above and let $f : \mathbb{R}^d \longrightarrow \mathbb{R}^p$ be a twice continuously differentiable function with $f = (f_1, \ldots, f_p)$. The process $Y_t = f(X_t)$ is given by the following stochastic differentials.

$$dY_t^k = \sum_{i=1}^{d} \frac{\partial f_k}{\partial x_i}(X_t)dX_t^i + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 f_k}{\partial x_i \partial x_j}(X_t)(dX_t^i \cdot dX_t^j).$$

The differentials are computed under the convention that $dW_t^i \cdot dW_t^j = \delta_{i,j}dt$.

Note that we may condensate the above equations in the following.

$$dY_t = f'(X_t)dX_t + \frac{1}{2}\sum_{k=1}^{p} tr\left[(dX_t)^T H_{f_k}(X_t)dX_t \right] e_k,$$

where f' is the Jacobian matrix of f, H_{f_k} is the Hessian matrix of f_k and $\{e_1, \ldots, e_k\}$ is the canonical basis of $\mathbb{R}^p$.

D.2 Stochastic differential equations

Consider a multidimensional Wiener process W and a random vector ξ, both of them defined over the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}^d$; suppose in addition that ξ is independent of the history $\sigma(\{W_t : t \geq 0\})$ of

the Wiener process. Let $\{\mathcal{F}_t\}_{t\geq 0}$ be the right-continuous and complete filtration generated by ξ and W. Moreover, let $a : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ and $b : \mathbb{R}^d \longrightarrow S_d^+$ be two continuous maps; where S_d^+ denotes de space of symmetric and positive semi-definite matrices. We are gonig to consider the stochastic differential equation

$$\begin{aligned} dX_t &= a(X_t)dt + b(X_t)dW_t, \\ X_0 &= \xi. \end{aligned} \tag{D.1}$$

The functions a and b are called, respectively, the drift and the dispersion coefficient of the SDE; the matrix $\sigma^2 = bb^T$ is known as the diffusion coefficient.

Definition D.2.1. A strong solution to equation (D.1) is a process X defined over $(\Omega, \mathcal{F}, \mathbb{P})$ and such that the following conditions hold.

1. X is $\mathcal{F}_t$-adapted and continuous.
2. $\mathbb{P}(X_0 = \xi) = 1$.
3. $\mathbb{P}\left(\int_0^t |a_i(X_s)| + |b_{i,j}(X_s)|^2 ds < \infty\right) = 1$ for all $t \geq 0$ and all $i, j \in \{1, \ldots, d\}$.
4. $\mathbb{P}\left(X_t = X_0 + \int_0^t a(X_s)ds + \int_0^t b(X_s)dW_s \quad \forall\ t \geq 0\right) = 1$

The following theorem is due to Itô.

Theorem D.2.2. Suppose that ξ is square integrable, namely $\mathbb{E}\,||\xi||^2 < \infty$, and that a and b satisfy the Lipschitz condition

$$||a(y) - a(x)|| + ||b(y) - b(x)|| \leq M\,||y - x|| \quad \forall\ x, y \in \mathbb{R}^d,$$

for some $M \geq 0$. There exists a strong solution X to equation (D.1), which is square integrable: for each $T \geq 0$ there exists a constant $C \geq 0$, which only depends on M and T, such that

$$\mathbb{E}\,||X_t||^2 \leq C(1 + \mathbb{E}\,||\xi||^2)e^{Ct} \quad \forall\ t \in [0, T].$$

Moreover, if Y is another strong solution then $X = Y$ with probability one.

Consider coefficients a and b satisfying the hypothesis of the last theorem. For each initial condition equation (D.1) gives us a unique strong solution. It is possible to see that this defines a Feller diffusion X whose infinitesimal generator coincides in $C^2(\mathbb{R}^d)$ with the second order differential operator L of equation (C.3), specifically

$$Lf = \sum_{i=1}^d a_i \frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^d \sigma_{i,j}^2 \frac{\partial^2 f}{\partial x_i \partial x_j} \quad \forall\ f \in C^2(\mathbb{R}^d).$$

Remark D.2.3. The above existence and uniqueness theorem may be extended to the time inhomogeneous case, where a and b depend on the time variable. Here we must additionally request that there exists some $K \geq 0$ such that

$$||a(t, x)||^2 + ||b(t, x)||^2 \leq K(1 + ||x||^2) \quad \forall\ x \in \mathbb{R}^d,\ t \geq 0.$$

D.3 Invariant measures and ergodicity

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider a process $(X_t)_{t\geq 0}$ defined on it, taking values in $\mathbb{R}^d$. We say that X is stationary if its finite-dimensional distributions are invariant under time shifts. Specifically, the random tuples

$$(X_{t_1+h}, \dots, X_{t_n+h}) \sim (X_{t_1}, \dots, X_{t_n})$$

have the same distribution for all times $0 \leq t_1 < \cdots < t_n$ and each shift $h \geq 0$. When X is the solution to an SDE, and in particular a Markov process, it is possible to give another characterization of stationarity.

Suppose now that X is a Markov process with initial distribution ν and transition function P. We know by Proposition C.2.4 that the finite-dimensional distributions of X are determined by ν and P, because

$$\mathbb{P}(X_{t_1} \in \Gamma_1, \dots, X_{t_n} \in \Gamma_n) = \int_{\mathbb{R}^d} \nu(dx_0) \int_{\Gamma_1} P_{t_1}(x_0, dx_1) \dots \int_{\Gamma_n} P_{t_n - t_{n-1}}(x_{n-1}, dx_n)$$

for all $0 \leq t_1 < \cdots < t_n$ and any choice of Borel sets $\Gamma_1, \dots, \Gamma_n$.

Definition D.3.1. A probability measure π, defined on the Borel subsets of $\mathbb{R}^d$, is an invariant measure for the Markov process X if it satisfies the condition

$$\int_{\mathbb{R}^d} P_t(x, \Gamma)\pi(dx) = \pi(\Gamma).$$

for each $t \geq 0$ and all Borel sets Γ.

If we let π be an invariant measure for X, then the above definition, together with the Chapman-Kolmogorov property, yield

$$\begin{aligned}\int_{\mathbb{R}^d} \pi(dx) \int_{\Gamma} P_{t+h}(x, dz) &= \int_{\mathbb{R}^d} \pi(dx) \int_{\mathbb{R}^d} P_h(x, dy) \int_{\Gamma} P_t(y, dz) \\ &= \int_{\mathbb{R}^d} \pi(dy) \int_{\Gamma} P_t(y, dz)\end{aligned}$$

for all $h, t \geq 0$ and all Borel sets Γ. It is now easy to check that X is stationary whenever its initial distribution is π. Conversely, if X is stationary and has initial distribution ν, then

$$\int_{\mathbb{R}^d} P_t(x, \Gamma)\nu(dx) = \mathbb{P}(X_t \in \Gamma) = \mathbb{P}(X_0 \in \Gamma) = \nu(\Gamma)$$

for all $t \geq 0$ and all Borel sets Γ, which means that ν is invariant. Therefore, X is stationary if and only if its initial distribution is an invariant measure.

In the sequel we give criteria for the existence, uniqueness and ergodicity of invariant measures, in the special case where X is the Feller diffusion associated to equation (D.1); the notion of ergodicity is defined below.

D.3.1 Foster-Lyapunov criteria

Consider the Feller diffusion X on $\mathbb{R}^d$ associated to equation (D.1), with drift coefficient a, dispersion coefficient b and diffusion coefficient $\sigma^2 = bb^T$. Recall that the infinitesimal generator of X coincides on $C^2(\mathbb{R}^d)$ with

$$Lf = \sum_{i=1}^{d} a_i \frac{\partial f}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{d} \sigma_{i,j}^2 \frac{\partial^2 f}{\partial x_i \partial x_j} \quad \forall\ f \in C^2(\mathbb{R}^d),$$

this is the second order differential operator of equation (C.3).

Suppose that the diffusion coefficient σ^2 is non-singular, in the sense that there exists $\alpha > 0$ such that

$$x^T \sigma^2(y) x \geq \alpha x^T x \quad \forall\ x, y \in \mathbb{R}^d.$$

Under this hypothesis we have the next criteria for proving the existence and uniqueness of invariant measures.

Theorem D.3.2. Suppose that there exist a non-negative function $V \in C^2(\mathbb{R}^d)$, called Foster-Lyapunov function, and some $r > 0$ such that

1. $LV(x) \leq -1$ for all $x \in \mathbb{R}^d$ such that $||x|| > r$.
2. $V(x) \to +\infty$ as $x \to \infty$.

Then X admits a unique invariant measure π.

Given a signed measure μ taking values in the Borel subsets of $\mathbb{R}^d$, and a measurable function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ with $f \geq 1$, we define

$$||\mu||_f = \sup_{|g| \leq f} \left| \int_{\mathbb{R}^d} g(x) d\mu(x) \right|.$$

Definition D.3.3. Assume that X admits a unique invariant measure π. Given a measurable function $f \geq 1$ we say that X is f-exponentially ergodic if there exist $\alpha \in (0, 1)$ and a non-negative function M such that

$$||P_t(x, \cdot) - \pi(\cdot)||_f \leq M(x)\alpha^t \quad \forall\ t \geq 0,\ x \in \mathbb{R}^d.$$

Note that exponential ergodicity implies that, for each $x \in \mathbb{R}^d$, the measures $P_t(x, \cdot)$ converge in distribution to π as $t \to +\infty$. Indeed, suppose that X has this property, and let α and M be as in the above definition. Each bounded and continuous $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ may be normalized to a function that is smaller than one at each $y \in \mathbb{R}^d$, and using this observation we see that

$$\lim_{t \to +\infty} \left| \mathbb{E}_x[g(X_t)] - \int_{\mathbb{R}^d} g(y) \pi(dy) \right| \leq \lim_{t \to +\infty} M(x)\alpha^t \sup_{y \in \mathbb{R}^d} |g(y)| = 0,$$

where $\mathbb{E}_x$ denotes the expectation with respect to probability measure defined by X when the initial distribution is the unit mass at x; see Section C.2.

Theorem D.3.4. Assume that X admits a unique invariant measure. Moreover, suppose that there exist $c > 0$, $d \in \mathbb{R}$ and a non-negative $V \in C^2(\mathbb{R}^d)$ such that

1. $LV(x) \leq -cV(x) + d$ for all $x \in \mathbb{R}^d$.
2. $V(x) \to +\infty$ as $x \to \infty$.

Then X is $(V+1)$-exponentially ergodic.

It is worth pointing out that a Foster-Lyapunov function that satisfies the conditions of the last theorem also satisfies those of Theorem D.3.2.

Remark D.3.5. As a final remark we note that it is possible to weaken the hypothesis regarding the non-singularity of σ^2. Roughly speaking, this hypothesis guarantees that the noise in the dispersion term of equation (D.1) spreads in all directions. In the singular case this may still happen with the help of the drift term.

Suppose that a and σ^2 are infinitely differentiable. Let $\{e_1, \ldots, e_d\}$ be the canonical basis of $\mathbb{R}^d$ and consider the smooth vector fields

$$Y_0(x) = \sum_{i=1}^{d} \left(a_i(x) - \sum_{j=1}^{d} \frac{\partial \sigma^2_{i,j}(x)}{\partial j} \right) e_i \quad \text{and} \quad Y_k(x) = \sum_{j=1}^{d} \sigma^2_{i,j}(x) e_j \quad \forall\, k \in \{1, \ldots, d\}.$$

The previous theorem holds if the Lie algebra $\mathcal{L}$, generated by $\{Y_0, \ldots, Y_d\}$, has dimension d. Note that this condition is easy to corroborate in the especial case where $a(x) = Ax$ and σ^2 is constant. Indeed, if we let J_k denote the Jacobian of Y_k, then the Lie bracket between Y_0 and Y_k is given by

$$[Y_0, Y_k] = J_k Y_0 - J_0 Y_k = -A\sigma^2_k \quad \forall\, k = \{1, \ldots, d\},$$

where σ^2_k denotes the k-th row of σ^2. Hence, it is enough to check that the vector space generated by $\{\sigma^2_k, A\sigma^2_k : k = 1, \ldots, d\}$ has dimension d; note that this is true in the non-singular case, where σ^2 has linearly independent rows.

Appendix E

Additional material

This appendix contains the proofs of three proposition that were used throughout this work. The first one is a refinement of the Leibniz rule, the second one concerns uniform differentiability and the third one is a result from harmonic analysis.

E.1 Refinement of the Leibniz rule

Proposition E.1.1. Consider a function $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ such that:

1. g is continuous in $\mathbb{R}^2$ except $\bigcup_{s\in I} \{(s,t) : t \in \mathbb{R}\}$, where I has measure zero.

2. $g(s, \cdot\,)$ is differentiable for all $s \in \mathbb{R}$ with $\frac{\partial g(\cdot,\cdot)}{\partial t}$ locally bounded.

Then, for almost every $t \in \mathbb{R}$, we have

$$\frac{\partial}{\partial t}\int_0^t g(s,t)ds = g(t,t) + \int_0^t \frac{\partial g(s,t)}{\partial t}ds.$$

Proof. The proof is based on [9, Theorem 2.27]. Given $h \in \mathbb{R}$ we have

$$\begin{aligned}\frac{1}{h}\left[\int_0^{t+h} g(s,t+h)ds - \int_0^t g(s,t)ds\right] &= \frac{1}{h}\int_t^{t+h} g(s,t)ds \\ &+ \int_t^{t+h} \frac{g(s,t+h)-g(s,t)}{h}ds \\ &+ \int_0^t \frac{g(s,t+h)-g(s,t)}{h}ds.\end{aligned}$$

The first term converges to $g(t,t)$ for all $t \notin I$ as $h \to 0$ by the fundamental theorem of calculus.

For the second term we first observe that, given a sequence $h_n \to 0$, we have

$$\frac{\partial g(\,\cdot\,,t)}{\partial t} = \lim_{n\to\infty} \frac{g(\,\cdot\,,t+h_n) - g(\,\cdot\,,t)}{h_n}.$$

Thus, the function on the left is measurable. Moreover, by the mean value theorem

$$\left|\frac{g(s,t+h_n)-g(s,t)}{h_n}\right| \leq \sup_{\zeta,\tau\in(a,b)}\left|\frac{\partial g(\zeta,\tau)}{\partial t}\right| < +\infty \quad \forall\ s\in(a,b);$$

where (a,b) can be chosen to be any finite interval containing $[0,t]$. Hence, since

$$\lim_{n\to\infty} \mathbb{1}_{[t,t+h_n]}(s)\frac{g(s,t+h_n)-g(s,t)}{h_n} = 0 \quad \forall\ s\neq t,$$

by the dominated convergence theorem we have

$$\lim_{n\to\infty}\int_t^{t+h_n}\frac{g(s,t+h_n)-g(s,t)}{h_n}ds = 0.$$

Similarly, for the third term, since

$$\lim_{n\to\infty}\frac{g(s,t+h_n)-g(s,t)}{h_n} = \frac{\partial g(s,t)}{\partial t} \quad \forall\ s\in\mathbb{R},$$

we have, again by dominated convergence, that

$$\lim_{n\to\infty}\int_0^t\frac{g(s,t+h_n)-g(s,t)}{h_n}ds = \int_0^t\frac{\partial g(s,t)}{\partial t}ds.$$

□

E.2 Uniform differentiability

Proposition E.2.1. Let $f : U\subset\mathbb{R}^d \longrightarrow \mathbb{R}^d$ be a continuously differentiable function, defined on the open set U, and consider a compact set $K\subset U$. For all $\varepsilon>0$, there exists $\delta>0$ with the following property: if the linear segment $[x,y]$ is contained in K and $||y-x||<\delta$, then

$$\frac{||f(y)-f(x)-f'(x)(y-x)||}{||y-x||} < \varepsilon;$$

where $f'(x)$ denotes the Jacobian matrix of f at the point x.

Proof. Suppose that $f=[f_1\ \cdots\ f_d]^T$. It is enough to prove for all $i\in\{1,\dots,d\}$ that, for all $\varepsilon>0$, there exists $\delta>0$, such that $[x,y]\subset K$, and $||y-x||<\delta$, imply

$$\frac{|f_i(y)-f_i(x)-\nabla f_i(x)(y-x)|}{||y-x||} < \varepsilon.$$

Choose some $i\in\{1,\dots,d\}$ and let $g=f_i$. Also, fix some $x,y\in K$ and define

$$\eta = \frac{y-x}{||y-x||}, \quad t = ||y-x||\,.$$

Assume that $[x,y]\subset K$, by the mean value theorem there exists $s\in(0,t)$ such that

$$\frac{g(y)-g(x)}{||y-x||} = \frac{g(x+t\eta)-g(x)}{t} = \nabla g(x+s\eta)\eta.$$

As a result, we have

$$\frac{|g(y) - g(x) - \nabla g(x)(y-x)|}{||y-x||} = |\nabla g(x+s\eta)\eta - \nabla g(x)\eta| \leq ||\nabla g(x+s\eta) - \nabla g(x)||.$$

Since g is continuously differentiable and K is compact, there exists $\delta > 0$ such that $u, v \in K$, and $||v-u|| < \delta$, imply

$$||\nabla g(v) - \nabla g(u)|| < \varepsilon.$$

If $||y-x|| < \delta$, then $||x+s\eta - x|| = s\,||\eta|| < \delta$, and we also know that $x + s\eta \in K$ because $[x, y] \subset K$, therefore

$$||\nabla g(x+s\eta) - \nabla g(x)|| < \varepsilon.$$

This completes the proof. □

E.3 A result from harmonic analysis

Proposition E.3.1. Let f be a locally integrable function such that

$$\int_{\mathbb{R}^d} f(x)\varphi(x)dx = 0 \quad \forall\ \varphi \in C_c^\infty(\mathbb{R}^d),$$

then $f = 0$ almost everywhere.

Proof. First, suppose that f has bounded support and consider some non-negative $\varphi \in C_c^\infty(\mathbb{R}^d)$ such that

$$\int_{\mathbb{R}^d} \varphi(x)dx = 1.$$

For instance, we could let φ be an adequate normalization of

$$\psi(x) = \begin{cases} e^{-\frac{1}{1-||x||^2}} & \text{if } ||x|| < 1, \\ 0 & \text{if } ||x|| \geq 1; \end{cases}$$

in this case $||\cdot|| = ||\cdot||_2$ denotes the usual euclidean norm. For each $\lambda > 0$ consider the function $\varphi_\lambda(x) = \lambda^d \varphi(\lambda x)$; note that $\varphi_\lambda \in C_c^\infty(\mathbb{R}^d)$. The family $\{\varphi_\lambda\}_{\lambda>0}$ constitutes what is called an approximate identity and has the property that $||g - g * \varphi_\lambda||_1 \to 0$ as $\lambda \to +\infty$ for each $g \in L^1(\mathbb{R}^d)$, where

$$g * \varphi_\lambda(x) = \int_{\mathbb{R}^d} g(x-y)\varphi_\lambda(y)dy = \int_{\mathbb{R}^d} g(y)\varphi_\lambda(x-y)dy.$$

Since f has bounded support, then its sign $g(x) = \mathbb{1}_{f(x)>0} - \mathbb{1}_{f(x)<0}$ belongs to $L^1(\mathbb{R}^d)$, and therefore we have

$$||g - g * \varphi_\lambda||_1 \to 0 \quad \text{as } \lambda \to +\infty.$$

This limit also holds almost everywhere with respect to the Lebesgue measure for a suitable subsequence; we refer the reader to [9, Theorem 2.30]. Specifically, there exists a sequence $(\lambda_n)_{n\geq 1}$ such that the functions $g_n = g * \varphi_{\lambda_n}$ converge to g almost

everywhere. Furthermore, note that $g_n \in C_c^\infty(\mathbb{R}^d)$ for all $n \geq 1$ and

$$|g_n(x)| \leq \int_{\mathbb{R}^d} |g(x-y)\varphi_{\lambda_n}(y)|dy \leq \int_{\mathbb{R}^d} \varphi_{\lambda_n}(y)dy = 1 \quad \forall\ x \in \mathbb{R}^d.$$

We may now write the following equation.

$$\begin{aligned}\int_{\mathbb{R}^d} |f(x)|dx &= \int_{\mathbb{R}^d} f(x)g(x)dx \\ &= \int_{\mathbb{R}^d} f(x)g_n(x)dx + \int_{\mathbb{R}^d} f(x)\left[g(x) - g_n(x)\right]dx.\end{aligned}$$

The first term on the right-hand side is zero because $g_n \in C_c^\infty(\mathbb{R}^d)$. Also, the integrand in the second term converges almost everywhere to zero and is dominated by $2|f|$. Hence, the second term on the right-hand side converges to zero as well by dominated convergence. This proves that $f = 0$ almost everywhere.

For the general case, pick any ball B centered at the origin and some $\xi \in C_c^\infty(\mathbb{R}^d)$ such that $\xi(x) > 0$ for all $x \in B$. For example, we could let $\xi = \psi * \mathbb{1}_B$. Now $f\xi$ has bounded support and

$$\int_{\mathbb{R}^d} \left[f(x)\xi(x)\right]\varphi(x)dx = \int_{\mathbb{R}^d} f(x)\left[\xi(x)\varphi(x)\right]dx = 0 \quad \forall\ \varphi \in C_c^\infty(\mathbb{R}^d).$$

Therefore, $f\xi = 0$ almost everywhere, and this implies that $f\mathbb{1}_B = 0$ almost everywhere, because $\xi(x) > 0$ for all $x \in B$. Since B is arbitrary, this proves that $f = 0$ almost everywhere. □

Bibliography

[1] C. Bianca and C. Dogbe, "On the existence and uniqueness of invariant measure for multidimensional diffusion processes." *Nonlinear Studies*, vol. 24, no. 3, 2017.

[2] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 1999.

[3] V. I. Bogachev, N. V. Krylov, and M. Röckner, "Elliptic and parabolic equations for measures," *Russian Mathematical Surveys*, vol. 64, no. 6, p. 973, 2009.

[4] A. N. Borodin, *Stochastic Processes*. Springer, 2017.

[5] A. Borovkov, "On limit laws for service processes in multi-channel systems," *Siberian Mathematical Journal*, vol. 8, no. 5, pp. 746–763, 1967.

[6] R. Darling, J. R. Norris *et al.*, "Differential equation approximations for markov chains," *Probability surveys*, vol. 5, pp. 37–79, 2008.

[7] A. B. Dieker, X. Gao *et al.*, "Positive recurrence of piecewise ornstein–uhlenbeck processes and common quadratic lyapunov functions," *The Annals of Applied Probability*, vol. 23, no. 4, pp. 1291–1317, 2013.

[8] S. N. Ethier and T. G. Kurtz, *Markov processes: characterization and convergence*. John Wiley & Sons, 2009, vol. 282.

[9] G. B. Folland, *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

[10] N. Gast and B. Gaujal, "Markov chains with discontinuous drifts have differential inclusion limits," *Performance Evaluation*, vol. 69, no. 12, pp. 623–642, 2012.

[11] D. Goldsztajn, A. Ferragut, and F. Paganini, "A feedback control approach to dynamic speed scaling in computing systems," in *51st Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2017.

[12] D. Goldsztajn, A. Ferragut, and F. Paganini, "Feedback control of server instances for right sizing in the cloud," in *56th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2018.

[13] D. Goldsztajn, A. Ferragut, F. Paganini, and M. Jonckheere, "Controlling the number of active instances in a cloud environment," *ACM SIGMETRICS Performance Evaluation Review*, vol. 45, no. 2, pp. 15–20, 2018.

[14] S. Halfin and W. Whitt, "Heavy-traffic limits for queues with many exponential servers," *Operations research*, vol. 29, no. 3, pp. 567–588, 1981.

[15] D. L. Iglehart, "Limiting diffusion approximations for the many server queue and the repairman problem," *Journal of Applied Probability*, vol. 2, no. 2, pp. 429–441, 1965.

[16] O. Kallenberg, *Foundations of modern probability*. Springer Science & Business Media, 2006.

[17] I. Karatzas and S. Shreve, *Brownian motion and stochastic calculus*. Springer Science & Business Media, 2012, vol. 113.

[18] F. Kelly, *Reversibility and Stochastic Networks*. John Wiley & Sons, 1979.

[19] L. Kleinrock, *Queueing Systems Vol: I, Theory*. John Wiley & Sons, Incorporated, 1976.

[20] T. G. Kurtz, "Solutions of ordinary differential equations as limits of pure jump markov processes," *Journal of applied Probability*, vol. 7, no. 1, pp. 49–58, 1970.

[21] T. G. Kurtz, "Limit theorems for sequences of jump markov processes," *J Appl Probab*, vol. 8, no. 2, pp. 344–356, 1971.

[22] T. G. Kurtz *et al.*, "Strong approximation theorems for density dependent markov chains," *Stochastic Processes and their Applications*, vol. 6, no. 3, pp. 223–240, 1978.

[23] M. Meyer, *Continuous stochastic calculus with applications to finance*. CRC Press, 2000.

[24] S. P. Meyn and R. L. Tweedie, "Stability of markovian processes ii: Continuous-time processes and sampled chains," *Advances in Applied Probability*, vol. 25, no. 3, pp. 487–517, 1993.

[25] S. P. Meyn and R. L. Tweedie, "Stability of markovian processes iii: Foster–lyapunov criteria for continuous-time processes," *Advances in Applied Probability*, vol. 25, no. 3, pp. 518–548, 1993.

[26] D. Mukherjee, S. Dhara, S. C. Borst, and J. S. van Leeuwaarden, "Optimal service elasticity in large-scale distributed systems," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 1, p. 25, 2017.

[27] D. Mukherjee and A. Stolyar, "Join-idle-queue with service elasticity: large-scale asymptotics of a non-monotone system," *arXiv preprint arXiv:1803.07689*, 2018.

[28] L. M. Nguyen and A. L. Stolyar, "A service system with randomly behaving on-demand agents," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1. ACM, 2016, pp. 365–366.

[29] J. R. Norris, *Markov chains*. Cambridge university press, 1997.

[30] B. Øksendal, *Stochastic differential equations*. Springer, 2003.

[31] D. Revuz and M. Yor, *Continuous martingales and Brownian motion*. Springer Science & Business Media, 2013, vol. 293.

[32] P. Robert, *Stochastic networks and queues*. Springer Science & Business Media, 2013.

[33] L. C. G. Rogers and D. Williams, *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*. Cambridge university press, 2000, vol. 2.

[34] M. van der Boor, S. C. Borst, J. S. van Leeuwaarden, and D. Mukherjee, "Scalable load balancing in networked systems: A survey of recent advances," *arXiv preprint arXiv:1806.05444*, 2018.

[35] A. Wierman, L. L. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems: Optimality and robustness," *Performance Evaluation*, vol. 69, no. 12, pp. 601–622, 2012.