Introduction to Bayesian Thinking

Gui Araujo

1 Bayesian Probability Framework

The Bayesian interpretation of probability assumes that the understanding of nature is always attached to a point of view, and that the picture of the world given by data is fundamentally uncertain.[7, 5] From that, probability is defined as a measure of uncertainty, from a point of view of the models themselves when trying to produce statements about nature. Under this view, other forms of interpretations, such as frequencies or propensities, are understood as models of probability; models of assessment and control of uncertainty.

Following that interpretation, we can understand the Bayesian framework as an extension to propositional deductive logic, where, instead of binary 0 or 1 true/false values of propositions, we have a $[0, 1]$ interval of possible values of certainty about a proposition. In this view, the objects of attention are propositions, statements about information we have, and we use the tools of probability spaces as a mathematical treatment for measuring the degree of belief of a given proposition. Probability is then an operator over statements outputting their plausibility, given a model. Thus, models and parameters are treated as subjects of propositions, and therefore have their values and definitions attached to probability statements. This is done in contrast with the standard approach of having random variables as arguments of probabilities, instead of uncertain propositions about fixed variables.

We can define a logical sum and a logical product operations over propositions in order to recursively combine them into compound propositions. We define A, B as a product operation, meaning A *and* B; it is true only if both A and B are true. And $A + B$ as a sum operation, meaning A *or* B; it is true when at least one of A or B is true. Then, for example, we can talk about the proposition C defined as $C = A + B$, A *or* B. We saw that this means that C is true if A is true, regardless of B, or if B is true, regardless of A; so C is only false if A is false and B is false, which is equivalent to say that $\overline{C} = \overline{A}, \overline{B}$, that reads as: not C is equal to not A and not B, by the use of a negation operator. Figure (1) shows the *and* and *or* operations as Venn diagrams. This reasoning gives us a tool to transform between sum and product operations through the use of the negation operation:

$$\overline{A, B} = \overline{A} + \overline{B}, \tag{1}$$

$$\overline{A + B} = \overline{A}, \overline{B}. \tag{2}$$

We can understand these as logical rules relating propositions (note that we can consider every equality between propositions also as a proposition, that is trivially true if it's a rule). Now, treating propositions as varying in degree of belief, the same as varying in plausibility in the interval $[0, 1]$, we need to define a good operator to rigorously convey the notion of plausibility. These are probability distributions.

1.1 Probability Rules

Standard probability distributions, the mathematical objects that model the concept of probability, are linked with the conceptual measure of uncertainty by being compatible with its goals. We express these goals in the form of three

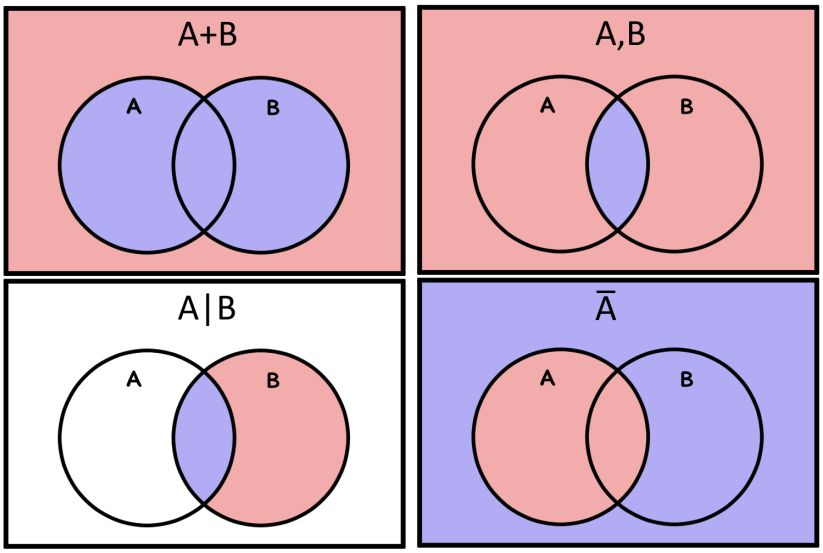

Figure 1: **Venn Diagrams**. The diagrams show a visual representation of the four most important logical expressions: *or* operation, *and* operation, conditional operation, and negation. Blue plus red regions represent the space of all possible outcomes, only blue shows the outcomes being considered by the composed propositions.

desiderata, asserting rules to be followed by the plausibility of every proposition:

I) Plausibility is represented by real numbers.

II) Plausibility must increase continuously and monotonically with the addition of information supporting the truth of propositions, as well as respect deductive limits. This is a desiderata of qualitative agreement with rational consideration of data.

III) Plausibility must be consistent: different ways of obtaining a result must give the same result; all given relevant information must be considered and equivalent propositions must be represented with equivalent plausibility.

With these three desiderata, we choose probability spaces as a good mathematical measure of plausibility. Thus, given a model, $P(A)$ is the plausibility of proposition A from that model, a number between 0 and 1. $P(A) = 0$ when A is certainly false and $P(A) = 1$ when A is certainly true. This identification

also provides the transformation of sum and product operations, given as the probability rules:

$$ProductRule : P(A, B) = P(A|B)P(B) = P(B|A)P(A) \tag{3}$$

$$SumRule : P(A + B) = P(A) + P(B) - P(A, B) \tag{4}$$

The proposition $A|B$ is the conditional proposition of A under B, read as A given B; it means the proposition A when we know that B is true (note that this does not imply temporal order between A and B, this is to stress that we are not talking about causal connections, but logical connections). With these rules, we build the negation of a proposition, $\overline{A}$, as

$$P(\overline{A}) = 1 - P(A). \tag{5}$$

Figure (1) shows Venn representations of a conditional proposition and of a negation of a proposition. A proposition and its negation form an exhaustive ($P(A + \overline{A}) = 1$) and mutually exclusive ($P(A, \overline{A}) = 0$) set, composed of two propositions. When N propositions are exhaustive and mutually exclusive, they are called a partition of the event space. Consider a partition represented as the set $\{E_i\}$, with $i = 0, 1, 2...N$. By applying the product and sum rules, we can write the probability of any proposition A using the conditionals over the $\{E_i\}$. This is called the law of total probability, and is useful for designing probability terms of conditioned propositions that we know something about:

$$P(A) = P((\sum_i E_i), A) = P(\sum_i E_i, A) = \sum_i P(E_i, A) = \sum_i P(A|E_i)P(E_i). \tag{6}$$

The first equality comes from the fact that $\{E_i\}$ are exhaustive, the second

equality is a distributive property of propositional sum. The third is the sum rule with $\{E_i\}$ being mutually exclusive, and then the product rule. As an example, suppose we have a set of dice, but with different numbers of sides. Then we randomly pick one to throw. We don't know directly the probability of any outcome, but we know the probability of each outcome given the die has n faces, so we can build the probability of an outcome using the law of total probability. In that case, the partition would be of propositions E_i = The die has i faces.

Another useful related operation is the marginalization of a joint probability, that is just a form of the law of total probability. Note what happens here,

$$P(A) = \sum_i P(E_i, A). \tag{7}$$

If we have the propositions A and E_i and their joint probability $P(E_i, A)$, we may use a partition built over E_i to remove E_i with a sum and obtain $P(A)$. This is called a marginalization of $P(E_i, A)$ over E_i. This is particularly relevant in the case where E_i asserts that some variable has a given value. Then we may consider as the partition the set with propositions for every value in that variable's domain; we marginalize over the domain of that variable. If that variable is continuous, we have $E_i = \{$The variable x is in the range between x_i and $x_i + dx\}$. Then we may write the marginalization process as

$$P(A) = \int dx_i P(E_i, A). \tag{8}$$

1.2 Bayes Equation

Statistical methods are mainly concerned with two major types of problems: 1) model selection, that uses data to establish criteria to choose between models for describing systems of interest; and 2) parameter estimation, that, given a

model, uses data to infer parameter values of the model. Both interests are centered around the Bayes equation, which is simply derived from the product rule of probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{9}$$

It has this quality of inverting the conditional, making it possible for us to update our knowledge about proposition A by the use of information acquired about B (because they're related, information is shared between them). This can translate into update of our theories in light of new data. Suppose we have a set of hypotheses $\{H_i\}$ and a proposition representing data, D. Then, with I representing our prior information on the matter, we have

$$P(H_i|D, I) = \frac{P(D|H_i, I)P(H_i|I)}{P(D|I)} = \frac{P(D|H_i)P(H_i|I)}{\sum_i P(D|H_i, I)P(H_i|I)}, \tag{10}$$

where the last equality is a law of total probability for D over the set of hypotheses (it's assumed that they form a partition, because they are naturally mutually exclusive, and, if they aren't exhaustive, it is rational to include the hypothesis that is the negation of the sum of every other, then making it exhaustive). This is the base of model selection and parameter estimation, as the set of hypotheses can represent alternate models or alternate versions of a model with different parameter values. This is also valid in continuous form, for a set of hypotheses parameterized by continuous values.

In this context of estimation by the use of data, we call $P(H_i|D, I)$ the posterior probability of H_i, that is the probability of H_i given that D is true, so given that we know the data. The term $P(D|H_i)$ is called likelihood of the data D over H_i, meaning the probability of the data D given that H_i is true. The term $P(H_i|I)$ is the prior probability of H_i, what we know before considering the data. The term in the denominator is of less importance and is usually

regarded as a normalization constant, with the estimation problem represented as

$$P(H_i|D, I) \propto P(D|H_i, I)P(H_i|I), \tag{11}$$

with the product of prior and likelihood acting as a kernel for the posterior, called the odds of that hypothesis in light of data. The prior information I is a formalization of our previous knowledge about the hypotheses. Normally, prior information becomes increasingly irrelevant as we accumulate more data.

The likelihood is where the probabilistic model of the system comes in. For example, we assume the validity of our dynamical model, that gives us the probabilities of the system being in every possibility of states, and with that we have probabilities of the system being where it was seen in the data. And by maximizing the likelihood, we can arrive at point estimations for the parameters, but a more robust treatment is made by obtaining the posterior distribution and summarizing it in all desired manners in order to obtain estimations concerning H_i. Model selection can be performed by calculating the odds ratio between hypotheses, that give the relative values of their posteriors.

This concludes our sketch of the general theory of the Bayesian framework considered in this work. We now turn to the Bayesian treatment of Markov jump processes, the basis of our stochastic modeling approach.

2 Stochastic Processes and the Markov Property

We start by defining a notion of a stochastic process. Our interest is in describing dynamical systems, which are systems changing over time. When we model these systems as stochastic, the model evolution is not completely known to us. The model dynamics is described by probabilistic trajectories over their states.

At first, we could think of propositions concerning the entirety of a dynamical process, but then we would restrict ourselves to probabilities about whole trajectories. So, following our interest in knowing probabilities dealing with each moment in time, we mostly consider propositions that are concerned with what is happening to the system at each instant. The definition of a stochastic process aims at pairing propositions with instants in time and chaining them in order to represent the whole dynamical process.

For us, stochastic process will be the set of propositions $\{X_{s,t}\}$, for $t \in T$, with T being the relevant set of time instants, for $s \in S$, with S being the relevant set of possible system states, where each $X_{s,t}$ is, in a general form, read as

$X_{s,t} = \{\text{The system is in state } s \text{ at time } t\}$.

Then, we may talk about the probability of the system being in state s at time t, $P(X_{s,t})$, or the probability of the system being at state s' at time t' if we know it's in state s at time t, $P(X_{s',t'}|X_{s,t})$. The models we study in this work are stochastic processes of a certain kind, they are Markovian models.

2.1 Markovian Stochastic Processes

Markovian processes are stochastic processes for which the probabilities associated with the system in future times depend only on knowledge about its current state. It means that the model doesn't hold any memory of previous states, and past knowledge has no bearing in its future. In practice, any conditioning on previous times actually is dependent only on the closest previous time. So, if $t_1 < t_2 < t_3$, we have that $P(X_{s_3,t_3}|X_{s_1,t_1}, X_{s_2,t_2}) = P(X_{s_3,t_3}|X_{s_2,t_2})$. This means that a transition between states is characterized only by the current state, regardless of the past. With that, we can talk about transition probabilities $P(X_{s',t_n}|X_{s,t_{n-1}})$, meaning a probability of the system moving to a state

s' at a time t' from the state s at a time t. Knowledge about some initial state and transition probabilities is sufficient to build the whole chain probability, just using the product rule together with the Markov property. For ease of notation, let's consider $X_{s_{t_n},t_n} = X_n$, with $t_n > t_m$ for $n > m$:

$$P(X_0, X_1, X_2) = P(X_2|X_0, X_1)P(X_0, X_1) = P(X_2|X_1)P(X_1|X_0)P(X_0). \quad (12)$$

Like that, we can build a chain for up to an arbitrary proposition X_n. This means that a Markovian model is completely characterized by an initial state and its transition probabilities (these transition probabilities are often presented as a transition matrix, and for a continuous time set T we often talk about *transition rates*).

The theory of Markovian processes spreads in four possibilities as we consider the nature of the states and times sets, S and T. Both can have continuous or discrete indexes for its elements. For discrete states and discrete times, we have the theory of discrete Markov chains, the system jumps over a network of states through discrete iterations. For continuous states and discrete times, we simply have the case of a continuous Markov chain, a case we'll briefly visit through the method of Markov chain Monte Carlo. For continuous states and continuous times, we have the Markov continuous stochastic processes, such as diffusion processes. Finally, for discrete states and continuous times, we have the Markov jump processes, the branch we are mainly concerned with in this work. Here, by a continuous passage of time, the system jumps between different states at random times.

Markov jump processes are governed by a Chapman-Kolmogorov differential equation that we call a master equation. The master equation is very often not solvable, and we can approximate it to continuous-states differential equations for continuous stochastic processes or even to deterministic differential equations

with a limit of infinite system.

2.2 Bayesian Derivation of a Master Equation for Markov Jump Processes

This section presents a derivation of the master equation using a Bayesian reasoning, and we emphasize our reliance on the simple probability rules defined in the previous section. We are interested in models that are Markovian processes of continuous time and discrete state space, Markov jump processes. Consider a system Γ of this type, defined by the following assumptions:

1. Γ exists in a discrete state space, with states that can be uniquely determined by a set of numbers, each describing a component of Γ (usually translated to integer count numbers of each type of component). So, if Γ is a system determined by two components, two species N_1 and N_2 with counts n_1 and n_2, then at a given time it's determined by the pair (n_1, n_2) contained in the set of possible states. We denote the state of the system with the vector $\boldsymbol{n}$ with dimension equal to the number of system's components. In the example, $\boldsymbol{n} = (n_1, n_2)$.

2. Γ evolves by changing states along a continuous passage of time. So, Γ has a continuous set of time instants and is a "jump process".

3. Γ obeys the Markovian property and we know the transition rates for the system. We'll rewrite them in terms of the transition probabilities. Also, the jumps to the many different states are independent events.

4. We can divide the time set into defined intervals dt for which we can consider $\mathcal{O}(dt^2)/dt \to 0$ for $dt \to 0$ and that we can assure transition rates to be approximately constant during dt.

So, if we know that Γ is in a state $\boldsymbol{n_1}$ at a time t_1, it can jump to any other state $\boldsymbol{n_2}$ at a posterior time t_2 with a probability $P(\Gamma_{\boldsymbol{n_2},t_2}/\Gamma_{\boldsymbol{n_1},t_1}, Z_{t_1,t_2}) = Tr(n_1, t_1 \to n_2, t_2)$, with $Z_{t_1,t_2} = \{$There are no jumps during the interval

$t_2 - t_1\}$ and $n_1 \neq n_2$. Since Γ is Markovian, the transition probability does not depend on states before t_1. Remember that $\Gamma_{\boldsymbol{n},t} = \{\Gamma$ is in state $\boldsymbol{n}$ at time $t\}$. In order to completely specify the system, we must connect the transition probabilities to the known transition rates. They are defined as follows:

$$Tr(n_1, t_1 \to n_2, t_2) = W_{\boldsymbol{n_1},t_1\to\boldsymbol{n_2},t_2}dt, \tag{13}$$

as long as $t_2 - t_1 = dt$. But we have problems. When the system jumps, this probability breaks; how many times can we expect it to jump during a time interval dt? Also, what is the probability of the system remaining in the same state after dt, $\overline{\sum_{\boldsymbol{n_j}} Tr(n, t \to n_j, t + dt)} = 1 - \sum_{\boldsymbol{n_j}} Tr(n, t \to n_j, t + dt)$? Can we know it?

2.2.1 Transitions

We are interested in the limit $dt \to 0$, so we can solve our problems by proving the following statement: During a passage of time dt starting at time t, the system can jump once, from state $\boldsymbol{n}$ to any different state $\boldsymbol{n_i}$ with probability $W_{\boldsymbol{n},t\to\boldsymbol{n_i},t+dt}dt$. Also, the system can jump more than once with probability $\mathcal{O}(dt^2)$ and remain in state $\boldsymbol{n}$ with probability $1-\sum_{n_i} W_{\boldsymbol{n},t\to\boldsymbol{n_2},t+dt}dt+\mathcal{O}(dt^2)$.

For that, consider the propositions, using the notation with implicit dependency of time $W_{n,n_j} = W_{\boldsymbol{n},t\to\boldsymbol{n_j},t+dt}$:

$K_k = \{$With Γ being in state $\boldsymbol{n}$ at time t, exactly $k > 0$ transitions occur during the next interval dt, k_j from $\boldsymbol{n}$ to $\boldsymbol{n_j} \neq \boldsymbol{n}$ with constant probability $W_{n,n_j}dt$ and the constraint $\sum_{n_j} k_j = k.\}$

With constant independent transitions, $P(K_k)$ follows a multinomial distribution with k trials and a number of possible outcomes equal to the number of possible states. One of its possible outcomes never happens in any trial,

representing the system jumping to nowhere in that trial, so:

$$P(K_k) = \sum_{\sum k_j = k} \frac{k!}{\prod_j k_j!} \prod_{n_j \neq n} (W_{n,n_j} dt)^{k_j} (1 - \sum_{n_j \neq n} W_{n,n_j} dt)^0 = (\sum_{\sum k_j = k} \frac{k!}{\prod_j k_j!} \prod_{n_j \neq n} W_{n,n_j}^{k_j}) dt^k. \tag{14}$$

We can see that this probability is proportional to dt^k, so we have $P(K_k) = \mathcal{O}(dt^k)$. In particular,

$$P(K_1) = \sum_{n_j \neq n} W_{n,n_j} dt. \tag{15}$$

For no transitions, we have the proposition K_0, defined as $K_0 = \overline{\sum_k K_k} = \overline{\sum_{\boldsymbol{n_j}} Tr(n, t \to n_j, t + dt)}$. Noting that the K_ks are mutually exclusive, using the sum rule, we have

$$P(K_0) = P(\overline{\sum_k K_k}) = 1 - P(\sum_k K_k) = 1 - \sum_k P(K_k) =$$

$$1 - P(K_1) - \sum_{k>1} P(K_k) = 1 - \sum_{n_j \neq n} W_{n,n_j} dt + \mathcal{O}(dt). \tag{16}$$

This ends our justification and solves our problems. We can now make sure that at most one transition occurs during dt in the limit.

2.2.2 Master Equation

Finally, we turn to the task of building the master equation. Let's give some easier names to our relevant propositions:

$X_0 = \Gamma_{\boldsymbol{n_0}, t_0} = \{\Gamma$ starts in an initial state $\boldsymbol{n_0}$ at time $t_0\}$.

$X = \Gamma_{\boldsymbol{n}, t} = \{\Gamma$ is in state $\boldsymbol{n}$ at time $t > t_0\}$.

And for each possible state $\boldsymbol{n}_i$:

$Y_i = \Gamma_{\boldsymbol{n_i}, t} = \{\Gamma$ is in state $\boldsymbol{n_i}$ at time $t' < t$ with $t' > t_0\}$.

The goal now is to assign a probability to proposition X using the Y_is. Let's look at the proposition $\sum_i Y_i$; it means Y_1, or Y_2, or Y_3, etc. It essentially means

that Γ is in any possible state at time t', and this is always true, the set $\{Y_i\}$ is exhaustive. Also note that Y_is are mutually exclusive, because at the same time Γ can only be in one state. So the set $\{Y_i\}$ is a partition of the event space at time t', a set of mutually exclusive events covering the whole space. Since the sum of Y_is is always true, and using product properties, we can write

$$X = (\sum_i Y_i), X = \sum_i X, Y_i. \tag{17}$$

Let's begin assigning probabilities to porpositions. Note that the products X, Y_i are also mutually exclusive, so we have, using the sum rule

$$P(X|X_0) = P(\sum_i X, Y_i|X_0) = \sum_i P(X, Y_i|X_0). \tag{18}$$

Now we use the product rule

$$P(X|X_0) = \sum_i P(X|Y_i, X_0)P(Y_i|X_0) \tag{19}$$

and then the Markovian property, that says $P(X|Y_i, X_0) = P(X|Y_i)$,

$$P(X|X_0) = \sum_i P(X|Y_i)P(Y_i|X_0). \tag{20}$$

See that all this is just the law of total probability applied to X with the partition $\{Y_i\}$. Now, why is it relevant to rewrite $P(X|X_0)$ in terms of the Y_is? It is because, with our specification of Γ, we have knowledge about local transition probabilities, but the known initial state X_0 may be as far as we wish from the arbitrary state X we want to describe. Using the Y_is as bridges, we can make t' "adjacent" to t and smuggle the known transition probabilities into our derivation. With adjacent meaning distant by an interval dt.

We need to specify a t' of Y_i that is adjacent to the t of X: $t' = t - dt$. If

this is true, we have the probabilities $P(X|Y_i)$ in terms of the transition rates. There are two cases; 1) $\boldsymbol{n_i} = \boldsymbol{n}$ and it means that no transitions occur, and 2) $\boldsymbol{n_i} \neq \boldsymbol{n}$ and it means that some transition with rate $W_{\boldsymbol{n_i},t-dt\to\boldsymbol{n},t}$ occurs. So we separate the sum in these two possibilities

$$P(X|X_0) = \sum_{\boldsymbol{n_i}\neq\boldsymbol{n}} P(X|Y_i)P(Y_i|X_0) + P(X|Y_n)P(Y_n|X_0), \tag{21}$$

with Y_n defined as Y_i for the case of $\boldsymbol{n_i} = \boldsymbol{n}$. The transition probabilities are, using the same implicit time-dependency notation as above,

$$P(X|Y_i) = W_{\boldsymbol{n_i},\boldsymbol{n}}dt + \mathcal{O}(dt^2) \tag{22}$$

because we are going from $\boldsymbol{n_i}$ to $\boldsymbol{n}$. The probability of no transition is

$$P(X|Y_n) = 1 - \sum_{\boldsymbol{n_i}\neq\boldsymbol{n}} W_{\boldsymbol{n},\boldsymbol{n_i}}dt + \mathcal{O}(dt^2) \tag{23}$$

because we are going from $\boldsymbol{n}$ to all other $\boldsymbol{n_i}$s. Note the exchange in the indexes of W. Putting more clearly, in case 1 the system is jumping from $\boldsymbol{n_i}$ to $\boldsymbol{n}$, and in case 2 the system already is in $\boldsymbol{n}$ and we consider the negation of it going to any other possible $\boldsymbol{n_i}$.

Inserting in the equation for $P(X|X_0)$, we have

$$P(X|X_0) = \sum_{\boldsymbol{n_i}\neq\boldsymbol{n}} \left(W_{\boldsymbol{n_i},\boldsymbol{n}}dt + \mathcal{O}(dt^2)\right) P(Y_i|X_0) +$$

$$\left(1 - \sum_{\boldsymbol{n_i}\neq\boldsymbol{n}} W_{\boldsymbol{n},\boldsymbol{n_i}}dt + \mathcal{O}(dt^2)\right) P(Y_n|X_0). \tag{24}$$

Just reorganizing the equation, we arrive at

$$\frac{P(X|X_0) - P(Y_n|X_0)}{dt} = \sum_{\boldsymbol{n_i} \neq \boldsymbol{n}} \left(W_{\mathbf{n_i},\boldsymbol{n}} P(Y_i|X_0) - W_{\boldsymbol{n},\boldsymbol{n_i}} P(Y_n|X_0)\right) + \frac{\mathcal{O}(dt^2)}{dt}. \tag{25}$$

Finally, we perform the limit $dt \to 0$. With this, the left side of the equation becomes the derivative of $P(X|X_0)$ in relation to time and $\frac{\mathcal{O}(dt^2)}{dt} \to 0$. $P(Y_n|X_0)$ on the right side becomes $P(X|X_0)$ as $t' \to t$ (note that the Y_is now represent Γ in time t with the limit imposing $t' \to t$). We have the Master Equation:

$$\frac{dP(X|X_0)}{dt} = \sum_{\boldsymbol{n_i} \neq \mathbf{n}} \left(W_{\boldsymbol{n_i},\boldsymbol{n}} P(Y_i|\{t' = t\}, X_0) - W_{\boldsymbol{n},\boldsymbol{n_i}} P(X|X_0)\right). \tag{26}$$

We can now change the probabilities to the more explicit distributions notation. The distribution that $P(X|X_0)$ follows has as variables the state vector $\boldsymbol{n}$ and the time t. If we define the probability of no transitions occurring as $W_{\boldsymbol{n},\boldsymbol{n}}$, we can sum over all states of Γ without altering the equation (note that the additional term $n_i = n$ ends up being zero). Calling the distribution $P(X|X_0) = \Pi(\boldsymbol{n}, t)$, we have

$$\frac{d\Pi(\boldsymbol{n}, t)}{dt} = \sum_{\boldsymbol{n_i}} \left(W_{\boldsymbol{n_i},\boldsymbol{n}} \Pi(\boldsymbol{n_i}, t) - W_{\boldsymbol{n},\boldsymbol{n_i}} \Pi(\boldsymbol{n}, t)\right),$$

$$\Pi(\boldsymbol{n}, t_0) = \delta(\boldsymbol{n}, \boldsymbol{n_0}). \tag{27}$$

Note that we can generalize the proposition X_0 into a set of propositions to mean that the state of Γ in t_0 is uncertain, with different probabilities of being in different states. We don't need to know the exact initial state for the equation to be valid. For systems with a finite number of states, we can even know nothing about the initial state, assigning to the set of X_0 a uniform probability distribution over the sates.

The solution of this equation gives the probability of proposition X happening once that X_0 happened, that means Γ has transitioned to state $\boldsymbol{n}$ after an arbitrary number of jumps during an arbitrary time interval $t - t_0$.

We can interpret the Master Equation in terms of gains and losses in probability; it means that the right side is viewed as a net gain in probability at time t, the first term being the gain from transitions into $\boldsymbol{n}$ and the second term being the loss from transitions away from $\boldsymbol{n}$.

The Master Equation is the differential form of the Chapman-Kolmogorov equation. In this work, we'll consider only time-independent transition rates (homogeneous Markovian systems), so the abridged notation, $W_{\boldsymbol{n_i},\boldsymbol{n}}$, will always be used.

3 Parameter Estimation

In this section, we'll connect model to data by building the parameter estimation process.[1] The input of the process is the data, measured from the physical systems of interest (this work is concerned specifically with reaction network systems and we'll test the estimation models using simulated data generated from stochastic simulations). The output of the process is a posterior probability function, the probability density of the estimated parameters under the model.

We start with the Bayes equation, that gives us the parameters' posterior distribution,

$$P(H_k|D, I) = \frac{P(D|H_k, I)P(H_k|I)}{P(D|I)}. \tag{28}$$

D is a proposition asserting the data, we'll specify it later. The set of hypotheses H_k will mean the following:

$H_k = \{\text{The model } m_k \text{ with parameter values } \boldsymbol{\theta_i} \text{ is true}\}$.

And we'll write $H_k = M_k, \Theta_i$, with these new propositions meaning:

$M_k = \{$The model m_k is true$\}$.

$\Theta_i = \{$The vector of parameters, $\boldsymbol{\theta}$, for the given model is between $\boldsymbol{\theta_i}$ and $\boldsymbol{\theta_i} + \boldsymbol{d\theta}\}$.

We are interested in parameter estimation and will work with a fixed model, so we may omit the proposition M_k as always true for our model. We also omit the data probability, since it doesn't involve Θ_i. Then, we'll work only with the estimation kernel, on the form

$$P(\Theta_i|D, I) \propto P(D|\Theta_i, I)P(\Theta_i|I). \tag{29}$$

There are two elements to deal with for the estimation, 1) the data likelihood given the parameter values, $P(D|\Theta_i, I)$, and 2) the parameter's prior information, $P(\Theta_i|I)$.

3.1 Priors

The choice of prior depends on everything we know about the model and the parameters. It can reflect the form of our likelihood, we may choose them to be conjugate, so the form of the posterior doesn't change by addition of new data to the estimation problem. It also depends on the nature of the parameters. We'll consider here only continuous parameters; then we may deal with two kinds of parameters: space parameters, that can be negative and depend on a choice of origin; and scale parameters, that are only positive and express our chosen units. The prior's domain must contain all known possible values for the parameters and can't contain values we know to be impossible (see that the posterior is zero for values outside this domain). It has to, at least in order of magnitude, represent our state of knowledge about probabilities for different parameter values. When we have no useful probabilistic information to use, we need to use non-informative priors; this can be tricky, because the

specifications of the problem may result in a nontrivial way to invoke uniformity over parameter spaces.

We may consider the posterior as a prior for inclusion of more data in the future, so the prior also loads information from possible previous measurements, and this also enables the notion of sequential data processing. Now, the specifics of prior choices is an experimental analysis and it depends on details of the problems, so we leave further developments to the applications when in need of them.

3.2 Data Likelihood

Before inputting the data into the likelihood, let's build the functional form of the likelihood. Consider a variable $\boldsymbol{y}$ to represent the data point at a time t. Then, assuming a continuous data variable just for convenience, the probability of having a data point valued $\boldsymbol{y}$ at this time is the same as the one for the proposition: $Y = \{\text{The data is in the range } [\boldsymbol{y}, \boldsymbol{y} + \boldsymbol{dy}] \text{ at time } t\}$. This analysis can be readily adapted to discrete variables. Now, consider what the model gives us. For each possible state of the model, $\boldsymbol{x}$, we have a probability for the system to be in that state at any time, given an initial state. The probability density for a continuous $\boldsymbol{x}$ is the same as the probability of the proposition $X_t|X_0'$ (remembering to consider an initial state), with: $X_t = \{\text{The state of the model is in the range } [\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{dx}] \text{ at the time } t\}$. If we consider a random error for each measurement, we have a variable $\boldsymbol{e}$ representing the possible values for the error, following a probability density equal to the one for the proposition: $E = \{\text{The measurement error is in the range } [\boldsymbol{e}, \boldsymbol{e} + \boldsymbol{de}]\}$. Now we can write the data $\boldsymbol{y}$ in terms of the variables representing the model and the error. If we define the variables $\boldsymbol{x}$ and $\boldsymbol{e}$ with the propositions $X_t|X_0'$ and E, we have that

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{e}. \tag{30}$$

This equation represents a measurement model. The function $f(\boldsymbol{x})$ is any transformation of the physical quantities that matches their relation to the measured quantity. The simplest form of the measurement model is one where we consider no measurement errors and the measurements are exactly the variables of our physical model, so $\boldsymbol{y} = \boldsymbol{x}$. From now on, we'll consider this simple relation between $\boldsymbol{y}$ and $\boldsymbol{x}$ but with measurement errors, so

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}. \tag{31}$$

This means, for example, that, if our physical model gives us molecular concentrations for $\boldsymbol{x}$, we are directly measuring those same molecular concentrations, but with with error e.

We want to compute the likelihood $P(Y|\Theta_i, I)$ as a function of $\boldsymbol{y}$, but in terms of what we know, that are the densities of $\boldsymbol{x}$ and $\boldsymbol{e}$. Here, we assume enough control over the measurement errors to know their probability densities.

Whenever we want to express a probability of a proposition in terms of other propositions, we may want to marginalize their joint probability, in a maneuver to include these propositions just to remove them again:

$$P(Y|\Theta_i, I) = \int\int \boldsymbol{dxde} P(Y, X_t, E|X'_0, \Theta_i, I), \tag{32}$$

We want a probability density on the variable $\boldsymbol{y}$ in terms of the parameters of the model and the measurement error, that's why we invoke the distributions for $\boldsymbol{x}$ and $\boldsymbol{e}$. With that distribution for $\boldsymbol{y}$, we'll later input the data value and get the probability of that data value given the model, i.e. the likelihood for that data. Using the product rule and considering that X and E are independent

(model independent of measurement errors),

$$P(Y|\Theta_i, I) = \iint \boldsymbol{dx}\boldsymbol{de} P(Y|X_t, E_j, X'_0, \Theta_i, I) P(X_t|X'_0, \Theta_i, I) P(E, \Theta_i, I). \quad (33)$$

Using the fact that $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$, we have

$$P(Y|X_t, E, X'_0, \Theta_i, I) = \delta(\boldsymbol{y} - \boldsymbol{x} - \boldsymbol{e}). \quad (34)$$

Putting $\delta(\boldsymbol{y} - \boldsymbol{x} - \boldsymbol{e})$ inside that integral has the effect of singling out the value of $\boldsymbol{e}$. Call the proposition:

$E_{y-x} = \{$The measurement errors are between $\boldsymbol{y}-\boldsymbol{x}$ and $(\boldsymbol{y}+\boldsymbol{dy})-(\boldsymbol{x}+\boldsymbol{dx})\}$.

Then, we end up with

$$P(Y|\Theta_i, I) = \int \boldsymbol{dx} P(X_t|X'_0, \Theta_i, I) P(E_{y-x}|I). \quad (35)$$

If we consider $P(X_t|X'_0, \Theta_i, I) = f_m(\boldsymbol{x}, \boldsymbol{\theta_i}, t)$, the model distribution over $\boldsymbol{x}$, and $P(E_{y-x}|I) = f_E(\boldsymbol{y} - \boldsymbol{x})$, the error distribution over $\boldsymbol{e}$,

$$P(Y|\Theta_i, I) = \int \boldsymbol{dx} f_m(\boldsymbol{x}, \boldsymbol{\theta_i}, t) f_E(\boldsymbol{y} - \boldsymbol{x}). \quad (36)$$

You may recognize this equation as a convolution integral. This relates to the fact that $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$, the sum of variables is computed as a convolution at the level of probabilities of propositions asserting those variables.

A simple and usual error model is the following: we assume the measurement error to be distributed as $\mathcal{N}(\boldsymbol{e}|0, \sigma_e^2 \boldsymbol{I})$, a multivariate normal distribution with mean zero and a known, constant, standard deviation over all measured variables and data points ($\boldsymbol{I}$ is the identity matrix of dimension equal to the system's dimension). With that, we have $f_E(\boldsymbol{y} - \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} - \boldsymbol{x}|0, \sigma_e^2 \boldsymbol{I})$. We now assume the simplest kind of model for the measurement process: a deterministic model.

In that case, $f_m(\boldsymbol{x}, \boldsymbol{\theta_i}) = \mu(\boldsymbol{x}, \boldsymbol{\theta_i})$ is a deterministic function of the data and the parameters. For example, in the case of a linear regression of a one dimensional model, we have $\mu(\boldsymbol{x}, \boldsymbol{\theta_i}) = \theta_1 x + \theta_2$. The result of the convolution will then be trivial

$$P(Y|\Theta_i, I) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}(\boldsymbol{\theta_i}, t), \sigma_e^2 \boldsymbol{I}). \tag{37}$$

By having the data points $(\boldsymbol{y}, t)$, we evaluate this likelihood in terms of the parameters $\boldsymbol{\theta_i}$. If we use a different value for the vector of parameters, say $\boldsymbol{\theta_{i'}}$, we obtain another value of the likelihood. The posterior probability density will then be a function of the parameters, given by the prior and likelihood functions.

Let's talk about the data proposition D in order to quantify the likelihood function using a data set obtained from a measurement operation. First, let's suppose that all variables specifying the state of the model are observed in a measurement, so a single measurement gives us a data vector $\boldsymbol{w}$ that is the same dimension as our physical model. This vector represents the observed quantities related to the variables of the physical model. Then we have for the data:

$D = \{$A measurement observed a set of d data points $\{([\boldsymbol{w_j}, \boldsymbol{w_j} + \boldsymbol{dw}], [t_j, t_j + dt])\}$ with $j = 0, 1, ..., d-1$, where $\boldsymbol{w_j}$ is a vector for the j-th measurement and t_j the measured time at that point.$\}$.

We have $D = \prod_{j=0}^{d-1} D_j$ for the D_j individual measurements. If we consider a string of measurements given a model that's Markovian, we have, using the product rule and the Markovian property

$$P(D|\Theta_i, I) = P(D_0|\Theta_i, I) \prod_{j=1}^{d-1} P(D_j|D_{j-1}, \Theta_i, I). \tag{38}$$

Each term $P(D_j|D_{j-1},\Theta_i,I)$ is, as we calculated, given by

$$P(D_j|D_{j-1},\Theta_i,I)=\int \boldsymbol{dx}P(X_t|X_0',D_{j-1},\Theta_i,I)P(E_{w_j-x}|I). \tag{39}$$

Here, $t=t_j-t_{j-1}$. How can that conditioned state, $\boldsymbol{x'}$, be chosen? The initial time and a possible initial conditioned state are given with the earlier measurement, D_{j-1}. That's the last observation we have of a system that we are modeling as Markovian, so let's use D_{j-1} from now on. In this scenario, we know that at the initial state the system is at $\boldsymbol{w_{j-1}}$ with an uncertainty given by $\boldsymbol{e}$. That's a random initial state for the model,

$$P(X_0'|D_{j-1},\Theta_i,I)=P(E_{w_{j-1}}|I)=f_E(\boldsymbol{w_{j-1}}). \tag{40}$$

We also need to compute the actual initial measurement of the whole data chain, $P(D_0|\Theta_i,I)$. What we do depends on the situation, we have to *model* the initial measurement. When the data is modelled by a normal distribution, it's natural to assume that $P(D_0|\Theta_i,I)$ is a normal distribution. Then, as a standard choice, we'll have the initial mean vector and variance matrix treated as parameters to be inferred, as parts of $\boldsymbol{\theta_i}$.

Then, following our deterministic model example, the likelihood for the data set D is given by

$$P(D|\Theta_i,I)=\mathcal{N}(\boldsymbol{w_0}|\boldsymbol{\theta_0},\boldsymbol{\theta_0'})\prod_{j=1}^{d-1}\mathcal{N}(\boldsymbol{w_j}|\boldsymbol{\mu_j}(\boldsymbol{\theta_i},\Delta t_j,\boldsymbol{w_{j-1}}),\sigma_e^2\boldsymbol{I}). \tag{41}$$

This will be a function of the data points $(\{\boldsymbol{w_j},t_j)\}$ and the model parameters $\boldsymbol{\theta_i}$. There we have, for this specific case, the whole likelihood function in terms of the model output and the data. The model output, in turn, is a function of the model parameters (what we want to estimate with the posterior).

3.3 Incomplete Measurements

Before, We assumed that all system's variables were observed. What happens if we have measurements only of a subset of the system's variables? It gets harder to estimate (even impossible for some kinds of parameters in certain systems). But the difference is that we also have to marginalize over the unobserved variables, to extract the subsection of the model that interacts with the data:

$$P(D_j|D_{j-1},\Theta_i,I) = \iint \boldsymbol{dxde} P(D_j|X_t,X_0',E,D_{j-1},\Theta_i,I)P(E|\Theta_i,I)\int \boldsymbol{du} P(X_t|X_0',D_{j-1},\Theta_i,I). \tag{42}$$

The vector $\boldsymbol{u}$ represents the unobserved variables, and the vectors coming from D have dimension equal to the number of observed quantities. The integral on $\boldsymbol{du}$ must be evaluated over the model's distribution. If it's a deterministic model (or even a possibly time-dependent normal distribution), the integral is the easiest possible, the result is just to ignore the unobserved components.

The posterior distribution is the final Bayesian result of the parameter estimation process. Once we have a posterior $P(\Theta_i|D,I)$ for the model parameters, we can summarize it by our decision in order to estimate the parameters in the ways suited to the systems we are analyzing.

4 Introduction to Markov Chain Monte Carlo

In the parameter estimation process, once the model is ready, we are in theory expected to integrate the kernel of the posterior for every set of values in the multidimensional parameter space; then, in order to extract information from the posterior we have to marginalize and calculate expectations through more integration on the posterior. Our task of estimating parameters transforms into a computational burden of integrating functions on a high dimensional space

and which normally feature a slim geometry of probability mass, making integration especially painful. For this reason, direct integration is virtually never a viable option in the Bayesian analysis. One of the main methods to determine probability distributions and widely used in Bayesian parameter inference is the Markov Chain Monte Carlo (MCMC).[8, 1, 4] Our goal is to use MCMC to calculate the parameter estimation process on the reaction network models. In this section, we will provide the intuition for this method, from the beginning.

4.1 Monte Carlo

A Monte Carlo method is one that in general transforms samples into integrals. This is built upon the law of large numbers, that basically shows us how to view uncertain events as certain events plus an approximation error.

We'll work out the intuitions through one dimensional continuous objects, but they can readily be generalized to more dimensions and discrete spaces. Suppose a data generating process $\{X_i^p\} = \{$The variable x modeled by the probability density $p(x)$ is in $[x_i, x_i + dx]\}$, with probability $P(X_i^p) = p(x_i)dx$. According to the law of large numbers, we can calculate the mean of any function $f(x)$ over a density $p(x)$ by using a set of N samples as the approximation

$$\langle f \rangle = \int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i), \tag{43}$$

an unbiased estimation with error $\mathcal{O}(N^{-1/2})$. As a particular well-known case, we have $\langle x \rangle \approx \sum_i x_i/N$, called the sample mean. But then, if we view $f(x)p(x)$ as a simple function of a real variable x, this is actually a method for calculating the definite integral of $f(x)p(x)$ over a support set through the sample mean. So the law of large numbers can act as a connection between samples of distributions and deterministic integrals. In particular, for a uniform density

over an interval of length L, we have $p(x) = L^{-1}$, and

$$\int_L f(x)dx \approx \frac{L}{N}\sum_{i=1}^{N} f(x_i). \tag{44}$$

In this case, we use the uniform samples as a sort of "mining" of function values that in the limit will equally distribute themselves around the function mean. And if we map the area under the curve of $f(x)$ into a rectangle by an area-preserving transformation, that rectangle would have a length of L and a height of $\langle f(x)\rangle$.

By using monte carlo integration, we can focus on just sampling the posterior. It is a much easier task than determining the posterior, marginalizing it, and calculating expectations.

4.2 Importance Sampling

The method in Eq. (44) presupposes that we draw samples from the distribution $p(x)$, but we may need or want to draw samples from another distribution $q(x)$, for example the standard case of drawing from uniform distributions in algorithms. Then, it would be useful if we could input the sampling from a different distribution $q(x)$ into calculations for $p(x)$. This can be done as the trick

$$\langle f\rangle = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i^q)\frac{p(x_i^q)}{q(x_i^q)}. \tag{45}$$

So it is the same as sampling the function $f(x)p(x)/q(x)$ from the $q(x)$ distribution. In this context, we can say that we are giving to each x_i^q an importance, or weight, of $p(x)/q(x)$ in order to calculate the mean of $f(x)$ under $p(x)$.

All this is, in principle, of great value for the parameter estimation process through the posterior distribution. With it, we may sample parameters from the posterior in order to calculate estimators for them, such as the mean, even

if we have to sample primarily from another distribution. But then we run into a problem: we can have at most the kernel of the posterior, not the entire density. So, we have the posterior represented by the density $p^*(x) = k(x)/Z$, where $k(x)$ is the kernel and $Z = \int_L k(x)dx$ is the unknown normalization factor of the posterior. But then, since Z is actually an integral, there is now a straightforward way to calculate it:

$$Z = \int_L k(x)dx \approx \frac{L}{N} \sum_i \frac{k(x_i^q)}{q(x_i^q)}. \tag{46}$$

Thus, by estimating Z itself, we can distribute importance (weight) to values of x in the interval according to an estimated density from the known kernel.

4.3 Rejection Sampling

But then, we notice that calculating from narrow distributions by sampling other densities like that may be an inefficient process. If $p(x)$ and $q(x)$ don't match, many samples x_i can have a negligible importance in relation to contributing to the probability mass, especially in high dimensional spaces. That's because the probability mass of kernels is usually concentrated in a narrow subset of the parameter space (called typical set), and it gets more concentrated for higher dimensions. This mismatch is the price we pay in order to sample from a distribution using another distribution. In an estimation task, if we could sample the x_i from the posterior itself, it would be a much more efficient sampling process, optimally efficient in this sense. A way to do this is to reject some x_i^q on the basis of their importance under the kernel. It makes the sample generation less computationally efficient to assure efficiency in the convergence of the integration. This transforms the sampling under $q(x)$ into a proposal of sampling, and a candidate sample is filtered under $p(x)$ (or the kernel). For us, a major advantage of this method is that we don't need to estimate Z, which

is a much more inefficient process. We'll see that the acceptance-rejection of x_i^q can be defined with only the kernel.

For each sampled x_i^q, we draw a uniform u in the interval $[0, 1]$, and we accept x_i^q if

$$u < \frac{k(x_i^q)/q(x_i^q)}{max[k(x)/q(x)]}, \tag{47}$$

intuitively meaning that, in order to be accepted, x_i^q must fall under the curve of k/q. Thus, we reject the sample if it falls off the curve of the kernel, in a region defined by the constant boundary $max[k(x)/q(x)]$ that ensures to encapsulate the whole curve of k/q. This boundary (and also q) is of course considered only from values inside the support of the kernel. Note that, by using a ratio as the filtering criterion, we don't need information of Z (it is only a scale on the kernel). In order to justify this, consider the propositions:

$A = \{\text{A value was accepted}\}$,

$X = \{\text{The sampled value is } x\}$.

Then, $P(X|A) = P(A|X)P(X)/P(A)$ has a density

$$p(x|A) = \frac{(k/(mq))q}{P(A)} = \frac{k(x)}{mP(A)}, \tag{48}$$

where we defined $m = max[k(x)/q(x)]$. $P(A)$, the probability of a proposal being accepted, irrespective of its value, can be calculated by marginalizing $P(X, A)$ over x:

$$P(A) = \int P(X, A)dx = \int P(X|A)P(X)dx = \int \frac{k(x)}{mq(x)}q(x)dx = \frac{Z}{m}. \tag{49}$$

Then, $p(x|A) = k(x)/Z$, and the accepted samples are distributed according to the desired density, in our case the posterior $p^*(x)$.

The most widespread picture of a monte carlo integration is done with rejection sampling. Instead of directly calculating the integral Eq. (44) from a

uniform sampling, the uniform sampling is used as a proposal. Then, the function $f(x)$ itself is used as a kernel for the rejection-acceptance step. The simple integral then equates with the monte carlo estimation of Z. The visualization of this process is one of dots accumulating both inside and outside the curve of $f(x)$; the dots falling inside the function are the accepted ones, and those falling outside are rejected.

4.4 Markov Chain

The task of determining a posterior distribution is one of finding its probability mass in the parameter space. We saw that a rejection sampling technique can assure that sampling will efficiently represent the posterior probability mass. But we just shifted the problem to the burden of proposing sample candidates. In high dimension parameter spaces, the probability mass will represent just a slim proportion of the space's volume. This means that a lot of proposals will get rejected if our choice of $Q(X)$ isn't already aligned with the kernel. Thus, we are still left with the pressing goal of electing an efficient proposal distribution, one that listens to the location of the posterior's probability mass.

The idea is to use the posterior's geometry in order to devise a criterion. In general, the probability mass is not scattered over the parameter space, but packed inside a specific typical set. We may guess that the typical set is concentrated around the mode, as is the case of the geometry of a one dimensional Gaussian distribution. But at higher dimensions, it non-trivially spreads away from the mode; because despite the importance of points is decreased, the volume of the typical set increases in regions away from the mode. Thus, the geometry of the high dimension posterior in general resembles a narrow band around the region of large importance. We must devise a sampling method that probes the parameter space for this set and then wanders over it with good

mixing.

This suggests that we correlate the sampling process, in an attempt to encode the goal of "getting closer" to the typical set once a sample falls far away from it. More formally, we want, given a sample, to distribute the next sample in a way that actively searches for probability mass. With that, $q(x)$ will shape itself according to the geometry of the posterior, granting a sufficiently high acceptance of proposals. For example, simply proposing samples that are nearby an accepted sample already does wonders in increasing the chances of acceptance, because we can expect that accepted samples are more probably located in good neighborhoods (the posterior mass is not scattered over the parameter space, but concentrated). In other words, if a sample is accepted, there is a higher chance that it is closer to the typical set than rejected samples, because the importance for acceptance is based on the kernel itself.

But if we want to correlate a sample with the previous sample, we want to make the sampling process into a Markovian chain. And since we want to lock it as being distributed as the posterior, it must be in equilibrium. Then the problem is reduced to the coordination of a proposal and an acceptance that result in both the equilibrium state of a Markov chain and the posterior distribution. In theory, no matter where the sampling process starts, it can converge to an equilibrium that mimics the sampling of the posterior. Since we now incur in the drawback of having correlated samples, we must ensure a good sampling mixture in order to use the process for estimations (ensure that the process is really able to capture the whole target distribution, and does not "get stuck" in certain regions).

Another problem to consider is that the acceptance process is not that well defined yet, because the determination of a quantity like $m = max(k/q)$ already is an optimization problem. The idea of probing for the typical set from a current

sample can also be used to address this and devise a local acceptance criterion.

This process of sampling from a Markov chain in order to calculate expectations from a desired target distribution is what is generally called a Markov chain Monte Carlo sampler (MCMC).

4.5 Metropolis-Hastings

The algorithm of Metropolis-Hastings is a MCMC sampler built on a property of reversible Markov chains, an equilibrium constraint called detailed balance. Consider the set of statements about a chain at equilibrium $\{X_i^{(t)}\} = \{\text{The state of the chain } x \text{ is in } [x_i, x_i + dx] \text{ at time } t\}$. Then, in detailed balance,

$$P(X_i^{(t-1)})P(X_j^{(t)}|X_i^{(t-1)}) = P(X_j^{(t-1)})P(X_i^{(t)}|X_j^{(t-1)}), \tag{50}$$

noting that $P(X_i^{(t-1)}) = P(X_i)$, because it is at the equilibrium. This is the same as saying that $P(X_i^{(t-1)}, X_j^{(t)}) = P(X_j^{(t-1)}, X_i^{(t)})$. Under detailed balance, the probability flux of the jump from i to j is the same as for the jump from j to i, so there is no net flux in the chain; the transitions are pairwise in equilibrium. When we define a particular chain through its transition probabilities, if we make sure that the chain satisfies detailed balance with the posterior, then if it is a proper posterior, that is the unique equilibrium distribution of the chain. Thus, the requirement is to choose transitions satisfying

$$\frac{P(X_j^{(t)}|X_i^{(t-1)})}{P(X_i^{(t)}|X_j^{(t-1)})} = \frac{k(x_i)}{k(x_j)}, \tag{51}$$

where $k(x)$ is the kernel of the posterior. The transition is the product of a proposal and an acceptance given proposal steps, so we must have

$$P(X_j^{(t)}|X_i^{(t-1)}) = q(x_i, x_j)P(A_{ij}), \tag{52}$$

where $q(x_i, x_j)$ is the sampling distribution, now dependent on both x_i and x_j, and $A_{ij} = \{\text{Given a proposal from } x_i \text{ to } x_j, \text{ the jump to } x_j \text{ is accepted}\}$. This results in

$$\frac{P(A_{ij})}{P(A_{ji})} = \frac{k(x_i)q(x_j, x_i)}{k(x_j)q(x_i, x_j)}. \tag{53}$$

If we chose

$$P(A_{ij}) = min\left(1, \frac{k(x_i)q(x_j, x_i)}{k(x_j)q(x_i, x_j)}\right), \tag{54}$$

then it is a valid distribution for which the condition is always satisfied.

The choice of a sampling proposal distribution $q(x_i, x_j)$ influences the speed of convergence of the chain. The particular Metropolis algorithm chooses it to be symmetrical (and making the acceptance independent of q), $q(x_i, x_j) = q(x_j, x_i)$, often a Gaussian $q(x_i, x_j) = \mathcal{N}(x_j|x_i, \sigma^2)$. In this case, the deviation σ regulates a step-size for proposals, that can't be too large so as to miss the regions of interest and cause a large rejection rate or too small so as to be slow on convergence and mixing.

For a multi-dimensional parameter space, there is also a choice involved in the jumps being sequential on each dimension or in form of a batch update (updating all dimensions at once is more efficient). The samples taken before convergence are discarded in the estimation process (the initial samples are called warm up), and parallel exploration with multiple chains is advisable. There are actually many details to address in the practical use of MCMC to carry out the estimation process and also the diagnosis and analysis processes following it. We then rely on a good software that can take care of much of the engineering bits.

4.6 STAN

We implement the MCMC method through the STAN statistical programming language.[3] It is a multi-interface language for custom Bayesian computation through advanced MCMC algorithms, written in C++. STAN targets a log transformation of the posterior distribution (needed for improved computation stability) and can run with two gradient-based MCMC methods for adaptive probing of parameter space. The Hamiltonian MCMC (HMCMC) and its variant No U-turn Sampler (NUTS).[2, 6] The motivation behind HMCMC methods is to interpret the posterior landscape as a potential energy, with a simulated Hamiltonian dynamics imprinting a momentum into the Markov chain, so the jump-size of the proposal adapts according to the gradient of the posterior; that is in contrast with the rigid Gaussian proposal. The NUTS variant implements strategies to improve the covering of the posterior by increasing the awareness of the chain based on the samples already visited; it is able to provide an automatic stop to the posterior's exploration, when the chain "sees" that it's enough.

STAN also contains many diagnosis, analysis, and visualization tools aiming at efficiently automate all tasks that are not related to the modeling aspect of the inference. It is possible to effortlessly check for convergence, mixing, and proportion of effective samples. There are also in-built transformations for constraining parameters and calculating the posterior in the log space.

5 Estimation Example: Lotka-Volterra

In order to properly illustrate the parameter estimation process, we consider the Lotka-Volterra model from Eq. (**??**). We generate data using the stochastic simulation algorithm on a medium-sized space of $\Omega = 100$: from a sample trajectory, we extract 30 measurements at random times for both preys and

predators.

We then use the deterministic dynamics as a statistical model of the data-generating process, thus the likelihood becomes a Gaussian having the model as the mean and the variance σ^2 as a proxy noise level also to be estimated. For simplicity, we use the generally well-suited exponential priors for all the parameters. The priors reflect the order of magnitude of the parameters, with means equal to one of 0.1, 1, or 10, depending on the parameter.

The statistical task is to estimate the parameters $\boldsymbol{\theta} = (w, \gamma, \delta, \mu, x_0, y_0, \sigma)$, where (x_0, y_0) is the initial state of preys and predators used to initialize the statistical model, and σ is the proxy standard deviation of the measurements. The initial state, for $t = 0$, was chosen as 200 preys and 100 predators, meaning densities of $(x_0 = 2, y_0 = 1)$. Figure (2) shows the sampled trajectory and measurements together with both the deterministic model and the mean estimated model, along with trace plots of the posterior. The table (1) compares the estimated values with the real parameter values used for data generation.

The estimation process runs with 4 chains. For all chains, the trace plots indicate the expected behavior of the jumps, a "fuzzy caterpillar" shape, of well-mixed exploration of the posterior. Note that, with just the noisy extracted measurements and the statistical model, we are able to, in theory, estimate the particular place of the multi-dimensional parameter space in which the system operates.

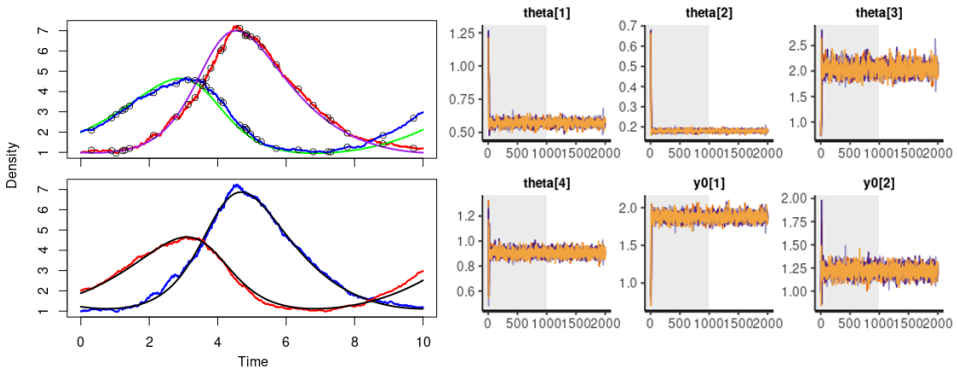

Figure 2: **Parameter estimation on the Lotka-Volterra model**. **Upper left:** Stochastic sample with 30 randomly extracted measurements for both preys and predators, compared with the deterministic model. **Lower left:** Same stochastic sample, compared with the deterministic curve generated with the mean estimated parameters. **Right:** Trace plots for posterior samples of the 4 network parameters and the initial state of the model. The gray region indicates the warm-up iterations

Table 1: **Estimated parameters.** The estimated posterior yields the mean values and standard deviations shown in the table, compared with the real parameter values. The standard deviation of the measurements σ has no real value because it is a proxy of the actual noise levels coming from the stochastic process.

Parameter	**Mean Estimation**	**Real Value**
w	0.57 ± 0.02	0.55
γ	0.18 ± 0.01	0.18
δ	2.04 ± 0.11	2.00
μ	0.91 ± 0.03	0.84
x_0	1.88 ± 0.05	2.00
y_0	1.22 ± 0.05	1.00
σ	0.15 ± 0.02	-

References

[1] Andrew Gelman et al. *Bayesian Data Analysis*. Third Edition. Chapman and Hall/CRC, 2013.

[2] Michael Betancourt. "A Conceptual Introduction to Hamiltonian Monte Carlo". In: (Jan. 2017). URL: http://arxiv.org/abs/1701.02434.

[3] Bob Carpenter et al. "¡i¿Stan¡/i¿ : A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.1 (2017). ISSN: 1548-7660. DOI: 10.18637/jss.v076.i01.

[4] Dani Gamerman and Hedibert Freitas Lopes. *Markov Chain Monte Carlo*. Tech. rep.

[5] P. C. (Philip Christopher) Gregory. *Bayesian logical data analysis for the physical sciences : a comparative approach with Mathematica support*. Cambridge University Press, 2005, p. 468. ISBN: 052184150X.

[6] Matthew D Hoffman and Andrew Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Tech. rep. 2014, pp. 1593–1623. URL: http://mcmc-jags.sourceforge.net.

[7] E. T. Jaynes. *Probability Theory: The Logic of Science*. Tech. rep.

[8] Steve Brooks et al. *Handbook of Markov Chain Monte Carlo*. Tech. rep.