VIETNAM NATIONAL UNIVERSITY - HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

MATHEMATICAL MODELING (CO2011)

Assignment

"The SIR model in COVID-19 prediction"

Advisors: Nguyen An Khuong
Nguyen Tien Thinh

Candidates: Nguyen Van A – 22102134
Tran Van B – 88471475
Le Thi C – 36811334
Pham Ngoc D – 97501334
Kieu Thi E – 12341334

Ho Chi Minh City, 07/2020

Contents

1 Introduction

1.1 Motivation

The first infected case of COVID-19 was reported in Wuhan (China) at the end of 2019. Up to now, the pandemic has spread around the world. Many nations have recently declared a state of emergency. In the United States, more than two million infections and more than one hundred thousand deaths have been reported on June 18th, 2020. Respectively, 25.9% and 26.24% of the total confirmed cases across the globe.

On the other hand, many governments have more or less imposed strict quarantine, isolation, and social distancing measures on their communities in order to prevent COVID-19 from new outbreaks. Nonetheless, new outbreaks might occur in the future unless the disease is studied carefully.

1.2 The SIR model

1.2.1 Introduction

The SIR (Susceptible - Infectious - Recovered) model is a basic compartmental model describing how individuals in different compartments in a population interact when there is an infectious disease. The model was first introduced in the early twentieth century (see [KM27]). The model is a system of three ordinary differential equations representing three compartments of the population: susceptible compartment, infectious compartment, and recovered compartment. In the model, we also assume that who recovered from the disease will be immune to it in the future and the total population does not change in time. The model is as follows.

$$\frac{dS}{dt} = -\frac{\beta}{N} IS, \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta}{N} IS - \gamma I, \tag{2}$$

$$\frac{dR}{dt} = \gamma I, \tag{3}$$

where at the time $t \geq t_0 \geq 0$, t_0 is the first time when an infection of the disease is reported,

- $S(t)$: The number of people who are susceptible to the disease;
- $I(t)$: The number of infected people;
- $R(t)$: The number of recovered people;
- $\beta(t)$: The contact rate between the susceptible compartment and the infectious compartment;
- $\gamma(t)$: The recovery rate when a person is infected;
- $N(t)$: Total population and it is defined as the sum of the three compartments.

$$N(t) := S(t) + I(t) + R(t). \tag{4}$$

- Equation (1) represents the decrease in time of the susceptible compartment. The decrease rate can be seen as the probability of the event that a susceptible individual is infected when he or she interacts with infected people.

- Equation (2) represents the change in time of the infectious compartment. It is obtained by subtracting the number of the new recovered people with the recovery rate γ from the number of the infected people and by adding the number of the new infected people to the number of the infected people;
- Equation (3) represents the increase in time of the recovered compartment. It is the number of the new recovered people with the recovery rate γ.

Example 1.1. Assume that a type of flu is spreading within a community. We also assume that

- The community is isolated, i.e., no one leaves or enters;
- The recovery time of an infected individual is exactly 2 weeks and it does not change in time;
- Who recovered from the flu will be immune to it in the future;
- A susceptible individual becomes an infected individual with the rate 0.2% (β/N). We also assume that the rate does not change in time.

The SIR model is

$$\frac{dS}{dt} = -0.002IS, \tag{5}$$
$$\frac{dI}{dt} = 0.002IS - 0.5I, \tag{6}$$
$$\frac{dR}{dt} = 0.5I \tag{7}$$

1.2.2 The Euler method in solving the SIR system

The Euler method is a first-order method that is used to solve ordinary differential equations. The method was first introduced by Leonhard Euler in Institutionum Calculi Integralis published in between 1768 and 1770.

Consider the ordinary differential equation

$$y' = f(t, y(t)). \tag{8}$$

The main idea of the Euler method is to approximate y by a sequence $\{y_n\}$ such that

$$y_{n+1} := y_n + f(t_n, y_n)\Delta t. \tag{9}$$

Here Δt is the step of the approximation and $f(t, y(t))$ is the slope of the tangent line to the curve y at the time t.

In general, we can consider the following system of ordinary differential equations

$$y_1' = f_1(t, y_1, \ldots, y_N), \tag{10}$$
$$\vdots$$
$$y_N' = f_N(t, y_1, \ldots, y_N), \tag{11}$$

where y_i is a real-value function depending on $t \geq 0$ and f_i is a real-value function depending on $t \geq 0$ and y_i for all $i \in \{1, \ldots, N\}$. Then we can apply the Euler method to approximate each y_i.

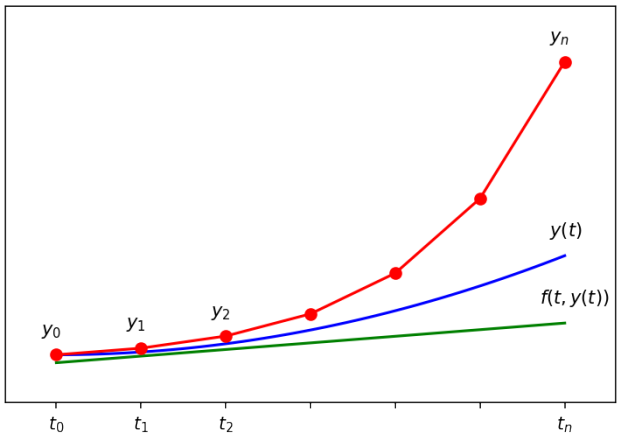

Figure 1: *The red curve is an approximation of y (blue) and the slope of the tangent line (green) of y is $f(t, y)$, i.e., $y' = f(t, y)$. The difference between y_n and y_{n-1} is exactly $f(t_{n-1}, y_{n-1})(t_n - t_{n-1})$*

Assume that at the beginning time, there are 800 susceptible people, 7 infected people, and there are no recovered people. By applying the Euler method, we can solve (5), (6), and (7) to find the number of people in each compartment (susceptible - infectious - recovered) in 8 weeks (2 months).

The approximation is as follows.

Week	Susceptible	Infectious	Recovered
0	800.000000	7.000000	0.000000
1	788.800000	14.700000	3.500000
2	765.609280	30.540720	10.850000
3	718.844763	62.034877	26.120360
4	629.657869	120.204332	57.137799
5	478.282662	211.477373	117.239965
6	275.990740	308.030609	222.978651
7	105.963549	324.042496	376.993955
8	37.290163	230.694633	539.015203

For reference, please read the books [ESG93; EG96].

1.2.3 Parameters estimation: β and γ

At the moment, quarantine measures can be seen as an effective way to reduce the number of infected cases. Hence, compartmental models such as SIR can be used to study the COVID-19 disease. Nonetheless, we will consider the coefficients β and γ as parameters depending on time since they depends on the ways that the governments and every individual reacts to the pandemic. The contact rate between the susceptible compartment and the infectious compartment then can change, and due to the fact that the number of COVID-19 patients and the quality of the services in the hospitals change day by day and are different in different regions, the recovery rate might be different at different periods of time.

The estimates for β and γ indeed depend on the confirmed cases of COVID-19, more specifically, the cummulative infected cases and the recovered cases. In this project, we will consider the Bayesian inference to estimate β and γ. Call

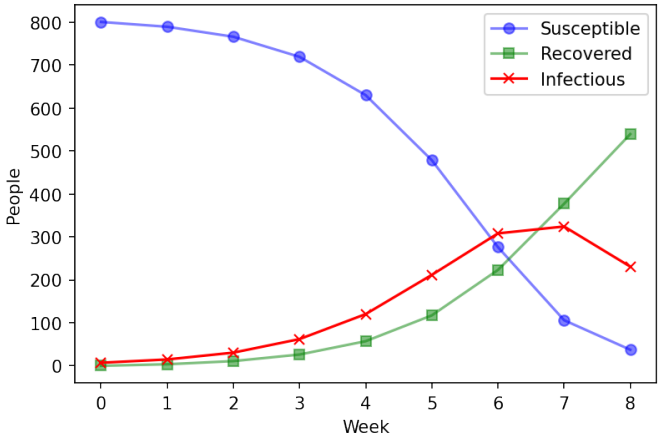

Figure 2: *The number of infected cases increases during the first 6 weeks and decreases after that*

- X: Random variable of infected cases and recovered cases at each time $t \geq t_0$;
- $\pi(\beta, \gamma|X)$: The posterior probability distribution of β and γ, provided confirmed infected and recovered cases;
- $\pi(X|\beta, \gamma)$: The likelihood of infected and recovered cases for given β and γ;
- $\pi(\beta, \gamma)$: The prior probability distribution of β and γ.

The Bayes theorem is stated as follows.

$$\pi(\beta, \gamma|X) \propto \pi(X|\beta, \gamma)\pi(\beta, \gamma). \tag{12}$$

It means that the posterior probability distribution of β and γ can be computed as the product of the likelihood of infected and recovered cases for given β and γ and the prior probability distribution of β and γ. The symbol $\propto$ means "be proportional to".

Example 1.2. Assume independent and identically distributed $X \sim \Gamma(\beta, \gamma)$. We also assume that the prior probability distributions of β and γ are Gamma distributions as follows.

$$\beta \sim \Gamma(\lambda_\beta, \nu_\beta) \tag{13}$$
$$\gamma \sim \Gamma(\lambda_\gamma, \nu_\gamma) \tag{14}$$

Hence, $\pi(\beta, \gamma|X)$ can be seen as the product of the following three probability distributions.

$$\pi(X|\beta, \gamma) = \prod_{i=1}^{n} f(X(t_i)|\beta, \gamma) = \prod_{i=1}^{n} \frac{\gamma^\beta}{\Gamma(\beta)} X(t_i)^{\beta-1} \exp\{-\gamma X(t_i)\}, \tag{15}$$

$$\pi(\beta) = \frac{{\nu_\beta}^{\lambda_\beta}}{\Gamma(\lambda_\beta)} \beta^{\lambda_\beta - 1} \exp\{-\nu_\beta \beta\}, \tag{16}$$

and

$$\pi(\gamma) = \frac{{\nu_\gamma}^{\lambda_\gamma}}{\Gamma(\lambda_\gamma)} \gamma^{\lambda_\gamma - 1} \exp\{-\nu_\gamma \gamma\}, \tag{17}$$

where n is the number of times we record the value of the random variable X and Γ is the Gamma function defined by the integral

$$\Gamma(y) = \int_0^\infty z^{y-1} \exp(-z) dz. \tag{18}$$

The estimates for β and γ play important roles. In fact, we are interested in the coefficient

$$R_0 := \frac{\beta}{\gamma}. \tag{19}$$

If $R_0 < 1$, there will be no new outbreak of the virus in the future since the contact rate β is strictly less than the recovery rate γ. If $R_0 > 1$, new outbreaks of the virus might occur in the future since the contact rate β is strictly larger than the recovery rate γ. Particularly, the mean of R_0 is the integral

$$E(R_0) = \int \pi(\beta, \gamma|X) R_0(\beta, \gamma) d(\beta, \gamma) \tag{20}$$

where X is the confirmed data of infected cases and recovered cases. This value can be estimated based on $\pi(\beta, \gamma|X)$. Indeed, we can use (12).

However, the integral (20) is nearly impossible to be computed directly. Instead, we can approximate it by

$$E(R_0) = \int \pi(\beta, \gamma|X) R_0(\beta, \gamma) d(\beta, \gamma) \propto \int \pi(X|\beta, \gamma) \pi(\beta, \gamma) R_0(\beta, \gamma) d(\beta, \gamma) \approx \sum_{i=1}^{m} \pi(X|\beta_i, \gamma_i) \frac{\beta_i}{\gamma_i}, \tag{21}$$

where (β_i, γ_i) is selected from the probability distribution $\pi(\beta, \gamma)$ and m is the size of the sample.

It amounts to saying that we can take m values of (β, γ) from the probability distribution $\pi(\beta, \gamma)$ and take the sum of $\pi(X|\beta, \gamma)\frac{\beta}{\gamma}$ over this sample. The Metropolis–Hastings algorithm can be then used to select such a sample of (β, γ).

The Metropolis–Hastings algorithm is well known as a Markov chain Monte Carlo method (see [Has70] and the citations therein). The algorithm has the following steps:

1. Initialize β_0 and γ_0.
2. Set $\beta := \beta_0$ and $\gamma := \gamma_0$.
3. Select β^* and γ^* randomly from an arbitrary probability distribution $p(\beta, \gamma)$.
4. If $p(\beta, \gamma)$ is symmetric, i.e., $p(\beta^*, \gamma^*|\beta, \gamma) = p(\beta, \gamma|\beta^*, \gamma^*)$, we set r as the probability to keep β^* and γ^*, which can be computed by the formula

   $$r := \min\left(1, \frac{\pi(\beta^*, \gamma^*)}{\pi(\beta, \gamma)}\right) \tag{22}$$

   and move to Step 6.
5. Unless, set

   $$r := \min\left(1, \frac{\pi(\beta^*, \gamma^*) p(\beta, \gamma|\beta^*, \gamma^*)}{\pi(\beta, \gamma) p(\beta^*, \gamma^*|\beta, \gamma)}\right) \tag{23}$$

   and move to Step 6.

6. Generate a value q randomly from a continuous uniform distribution $U(0,1)$.
7. If $q < r$, set $\beta_{i+1} := \beta^*$ and $\gamma_{i+1} := \gamma^*$, and move to Step 9.
8. Otherwise, set $\beta_{i+1} := \beta_i$ and $\gamma_{i+1} := \gamma_i$, and move to Step 9.
9. Repeat Step 2 with $\beta := \beta_i$ and $\gamma := \gamma_i$ until we get m elements.

For more details, please take a look at the book [Bro+11].

1.3 Using Machine learning models

1.3.1 Introduction

To have a better comprehension of β and γ, Machine Learning models are considered. The significance here is the make use of the computational ability of machines in order to recognize more complex patterns in the data. Machine Learning models can be seen as neural networks, in which neurons connect to each other from layer to layer. That helps the model to look at deeper inside the data and extract important features. The original idea of Machine Learning indeed comes from Statistics. It was first introduced by Marvin Minsky and Dean Edmonds in 1951 with the help of the computers in the laboratory.

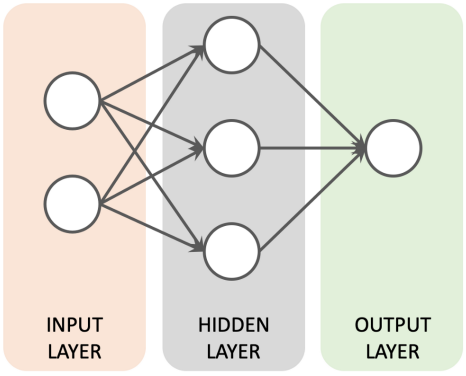

Figure 3: *A simple Machine Learning model with one hidden layer - A neural network*

1.3.2 Some difficulties in Machine Learning

The core of Machine Learning is an optimization problem. The model will find a minimizer for the so-called loss function, which is simply a function that measures the difference between predicted data and real data. A better loss function thus gives better prediction. However, depending on the data that we got, the design of such a loss function is quite challenged in general. Good minimization techniques are also important issues. There are also many other problems in Machine Learning you might be interested in.

1.3.3 Machine learning models in COVID-19 prediction

An example of the use of Machine Learning models in the study of COVID-19 is as follows. Firstly, an SIR or SIRD model with given initial values of β (the contact rate), γ (the recovery rate), and even μ (the death rate if an SIRD model is considered) is generated. Using that system

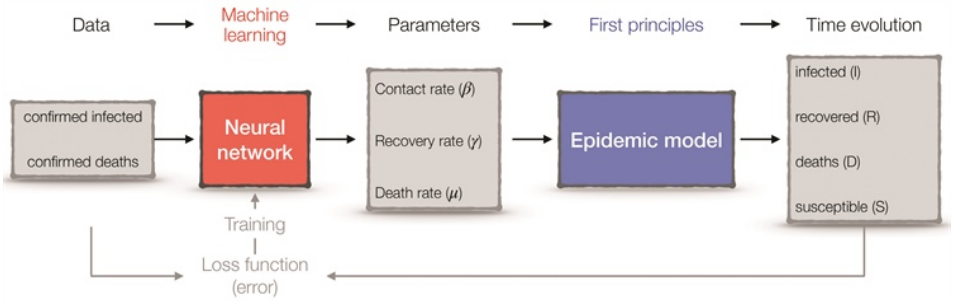

Figure 4: *The use of Machine Learning model proposed by Luca Magri and Nguyen Anh Khoa Doan in [MD20]*

of ordinary differential equations, we predict the number of infected people $I(t)$, the number of recovered people $R(t)$ as well as the number of deaths $D(t)$ at the time $t \geq t_0$ by applying, for instance, the Euler method. The predicted values and the confirmed cases are then fed into a neural network in order to adjust the recovery rate γ, the contact rate β as well as the death rate μ if we consider an SIRD model based on the minimization of the loss function. For more details, please take a look at the original paper [MD20]. Minimization techniques can be found in [SL04] and any Machine Learning and Deep Learning textbooks.

1.4 COVID-19 data

In this framework, we consider the confirmed data collected from all over the world and located in https://github.com/CSSEGISandData/COVID-19.The time-series data can be taken from the directory "COVID-19/csse_covid_19_data/csse_covid_19_time_series/".

2 Instructions and requirements

2.1 Instructions

Read carefully the models in reference documents [GFH13, Example 5 on page 50 and Example 3 on page 564] and [JKG20] (Vietnamese version - `https://tinyurl.com/y8bdl3hn`).

Each working group has from 3-5 students.

2.2 Requirements

- Due date: 28/7/2020. The answer of each question must be well written and correct.
- The report must follow **the sample layout** using LaTeX.
- The final submission must include **a log file**, which has the following information: the working tasks of each member, the discussion among the members, the details about the working process,...

2.3 Submission

- Only submissions through BK-eLearning will be accepted. Please compress everything (file .tex, file .py,...) as a single file named "*Assignment-CO2011-MT192-Student ID num-*

bers.zip". Go to the Assignment section on BK-eLearning and submit the compressed file.

- Noting that **only your group leader** will submit the file.

3 Exercises

Answer the following questions:

Exercise 1 (Compulsory)**.** Give a very detailed introduction and construction of the SIR model (for both discrete and continuous models) or its extensions including some specific examples of the model(s).

Exercise 2 (Compulsory)**.** Write a program to find an approximate solution to the SIR system or its extensions using the Euler method or its extensions. The input parameters consist of the time t, the contact rate β, the recovery rate γ, and the initial conditions of the SIR model: the total number of the population, the number of the infected people $I(t_0)$, and the number of the recovered people $R(t_0)$ where t_0 is the first time an infected case is reported. The program should return the number of infected cases $I(t)$ and the number of recovered cases $R(t)$ at the input time $t \geq t_0$. Implement your program with some specific values of the input parameters. If you are considering an extension of the SIR model, for instance, the SIRD model, the output of the program will be $I(t)$, $R(t)$, and $D(t)$, which is the number of deaths at the time $t \geq t_0$. The deaths are due to the disease and are not caused by any other reasons, for instance, the natural deaths. Plot the approximate solution. Include the details in your report.

Exercise 3 (Compulsory)**.** Write a program to perform the Metropolis–Hastings algorithm to take a sample from the probability distribution $\pi(\beta, \gamma)$. The sampler should return a sample of β and γ that has the probability distribution $\pi(\beta, \gamma)$. Plot the trace of the values of β and γ to describe the sampling process. Plot the sampling process. Include the details in your report.

Exercise 4 (Compulsory)**.** Select a region of nations, estimate the mean of R_0 given by (20) by using the Metropolis–Hastings sampler in Exercise 3 and the COVID-19 data of this region. Analyze the effect of the social distancing and the government policies on R_0. Cite all the reference documents. Please consider [JKG20] as a reference. A sample code can be found in [LM05] (written in R). Include the details in your report.

Exercise 5 (Compulsory for students of the honors program, optional for the others)**.** Consider the model and the loss function given in the article [MD20], train the model with the COVID-19 data in the selected region of nations. Analyze the result and visualize it by plotting. Include the details in your report.

4 Evaluation and dishonesty treatment

4.1 Evaluation

Submission will be evaluated as follows.

4.2 Dishonesty treatment

The assignment must be done by all of the students in each group. The following will be considered as cheating if

Item	Score (%)
Well-written and well-implemented codes	30%
Good and logical analysis following the question targets	30%
Correct diagrams and plots	20%
Having an appropriate and correct background knowledge section	15%
Well-written report	5%

- There are two or more reports with similar content (similarity in background knowledge section is the worst case). In this case, ALL OF the similar reports will be considered as cheating. Hence, students in each group should protect the report of the group from the others.
- There is a student in a group who does not understand anything about his/her working tasks. Students are allowed to use any documents for reference, but must be sure that you understand everything that you wrote in the report.

Cheating will be treated by the university stipulates.

References

[Bro+11] Steve Brooks et al. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[EG96] Hairer Ernst and Wanner Gerhard. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer, 1996.

[ESG93] Hairer Ernst, P. Nørsett Syvert, and Wanner Gerhard. *Solving Ordinary Differential Equations I: Nonstiff Prolemns*. Springer, 1993.

[GFH13] Frank Giordano, William P Fox, and Steven Horton. *A first course in mathematical modeling*. Nelson Education, 2013.

[Has70] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57 (1) (1970), pp. 97–109.

[JKG20] T. Wu Joseph, Leung Kathy, and Leung Gabriel. "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study". In: 395 (2020).

[KM27] William Ogilvy Kermack and A. G. McKendrick. "A contribution to the mathematical theory of epidemics". In: *Proc. R. Soc. Lond. A* 115 (1927), pp. 700–721.

[LM05] S. T. Ho Lam and A. Suchard Marc. *Simple MCMC under SIR*. 2005. URL: https://cran.r-project.org/web/packages/MultiBD/vignettes/SIR-MCMC.pdf.

[MD20] Luca Magri and Nguyen Anh Khoa Doan. "First-principles Machine Learning for COVID-19 Modeling". In: *arXiv preprint arXiv:2004.09478* (2020).

[SL04] Boyd Stephan and Vandenberghe Lieven. *Convex Optimization*. Cambridge University Press, 2004.