Probabilistic Graphical Models for COVID-19

Abstract

This project offers an exposition of COVID-19 modeling techniques based on the ideas and problem setup highlighted in (1).We define a generative model corresponding to our intuition about epidemiological modeling using the probabilistic programming framework `Pyro` (5) and apply probabilistic inference to draw insights into controlling the COVID-19 pandemic through interventions. In particular, we estimate the confidence intervals for the outbreak parameters to ensure that a predetermined goal is achieved. **We are not epidemiologists**; the sole aim of this study is to serve as a guide to generative modeling, not to draw inference about real-world impact of policy-making for COVID-19.

1. Introduction

COVID-19 modeling has recently been explored in light of the need to guide and support policy decisions (1) (2). This project provides to data scientists insight on how to go about building a scientifically accurate mathematical model for the spread of disease. This can be done by exploring the impact of policy decisions on synthetically generated data, or on distributions conditioned on the real-world. We hope to accurately obtain estimates for the proportion of the infected population at a given time, the infectiousness of the disease, predicted number of deaths, and other variables crucial to understanding an epidemic. Following this, we demonstrate the impact of performing interventions. Our project expands simulation-based inference using probabilistic programming as has been shown for physics-based simulators in (9)

2. Problem Setup

Disease modeling can be done at the population-level by assuming a global mathematical model that abstracts away individual level effects. A compartmental model is one such model that divides the population of individuals into different states or compartments at each point in time. The general idea is that individuals in a population are interacting with each other and as a result of this interaction they may transition from one compartment to another with a particular transition probability. Since all individuals are assumed to be equal, this transition need only be represented by a global transition probability between every set of subsequent compartments which encapsulates the notion of 'spread' of disease through a population through individual-level compartmental transitions. It helps that in practice, in addition to the number of deaths, the fraction of total population in the infected state/compartment at a particular time is also what we typically utilise in public awareness campaigns, which can be obtained directly from a compartmental model.

The simplest such compartmental model splits the population into three group, known as *Susceptible-Infected-Recovered* (SIR).

3. SIR Model Dynamics

It is important to take a look at the SIR model dynamics since they represent the essence of all the other compartmental models we explore. These dynamics are encapsulated in the following differential equations representing the mathematical model for SIR, but there are certain operating assumptions that need to be highlighted first.

- We work with the assumption that the total population is constant, ignoring new births and deaths for the time being, in this simplistic setup for epidemiological modeling of COVID-19.
- The rate of change of the susceptible population S at time t depends on the fraction of infected population at that time $i(t)$ since the number of infectious contacts relies on the fraction $i(t)$ of infected people in the population times the total number of contacts times the transmission probability–defined as the probability that an infectious contact may result in a spread of the disease.
- The transition from state I to R is fairly straightforward and obtained directly from the rate of recovery of the infected population.
- Consider model parameters β and γ where β represents the transmission probability of the disease and γ represents the recovery rate. The model infectiousness or reproduction number R_0 is defined as $\frac{\beta}{\gamma}$.
- These dynamics are encapsulated in the following differential equations representing the mathematical model for SIR.
- Note that we should formally only use $S(t)$, $I(t)$ and $R(t)$ but we sometimes omit it for simplicity.

$$s(t) + i(t) + r(t) = 1 \tag{1}$$
$$\frac{dS}{dt} = -\beta S(t) i(t) \tag{2}$$
$$\frac{dI}{dt} = \beta S(t) i(t) - \gamma I(t) \tag{3}$$
$$\frac{dR}{dt} = \gamma I(t) \tag{4}$$
$$\tag{5}$$

3.1 Advanced Compartmental Models

The SIR Model is a basic example of a compartmental model; we consider models with a higher number of compartments, more complex model dynamics, and ultimately, closer to reality by making weaker assumptions in modeling the data. One example of such a model is the Susceptible-Exposed-Infected-Recovered (SEIR) model which incorporates one additional compartment representing the fraction of population that has been exposed to

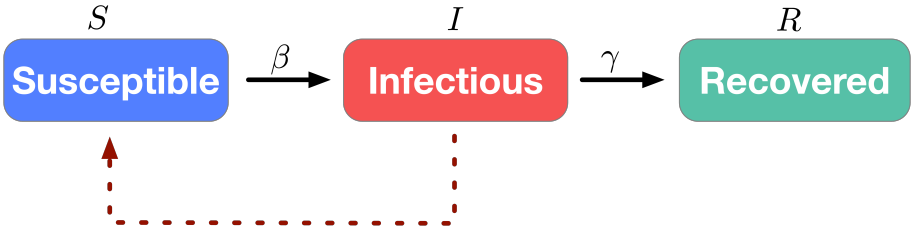


Figure 1: The Susceptible-Infected-Recovered (SIR) Model (by UCLA ML)

the disease. The claim behind this model is that infection logically follows from the exposure to a disease so it would be more realistic to model a separate set of people that may have been exposed to it.

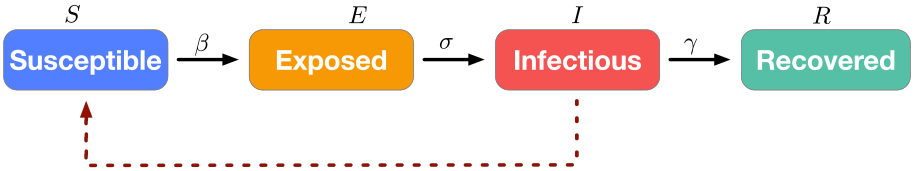


Figure 2: The Susceptible-Exposed-Infected-Recovered (SIR) Model (by UCLA ML)

4. Technical Details

We show the application of compartmental (SIR, SEIR) models to the spread of COVID-19 among a population. We introduce a policy intervention that allows us to search for minimally invasive global policies that can help us limit the spread of infectious diseases.

We start by placing priors on the non-controllable disease parameters as expressed in the paper (1) with updates where possible, based on recent literature on COVID-19. The goal is to infer the level of intervention required ('u', in the paper) such that the infectiousness parameters (dictating disease spread) are reduced to constrain the spread to lie within a predetermined level (e.g. 10% of the total population).

It is important to note the divergence in research objectives. Their goal is to demonstrate that an existing agent-based simulator for influenza written in C++ can be coupled with a Python-based Probabilistic Programming framework (Pyprob). Our goal is more in line with the project guidelines for 1005, which is to present their work as an end-to-end model in `Pyro` to replicate the core contributions of their study in a more accessible framework with better end-user support. We produce a project report supported by well-documented, modular code that can be offered as a tutorial contribution for the `Pyro` Epidemiology Docs.

4.1 Motivation for Probabilistic Programming

Probabilistic programming is a natural candidate among the tools we considered for this problem. We would like to define a model a time-series that encodes our assumptions about the data-generating process, fit that model to the data, and use the posterior distributions

to make predictions about future time-steps. `Pyro` is a good fit because of the following reasons:

- Comprises of thin wrappers around `PyTorch` distributions allowing us to write complex generative models interweaving stochastic and deterministic control flow while still within a familiar and popular ML framework.
- Offers general-purpose inference algorithms out of the box that allows us to easily frame our question and focus on experiments.
- Recently included support to develop compartmental models albeit with some limitations on performing inference in time-series setting so we want to test how easy/hard it is to perform a similar academic experiment in `Pyro`.

4.2 Approximate Inference in Probabilistic Programming Languages

Probabilistic inference gives us the ability to interpret and analyse complex graphical models. Specifically, it provides us with the tools to understand certain relationships in such models and thereby determine the flow of information in directed and undirected graphical models. Our goal is to infer the structure of latent variables in a class of probabilistic graphical models such that we update our prior beliefs about them. This allows us to reason about them in the real world. However, for a large class of models, it is intractable to perform exact inference in a general setting. This necessitates assumptions of factorization and conditional independence–which can be informally be described as the presence of structure in graphical models. This structureis designed to allow us to work around the intractability that arises in complex datasets with a large number latent variables.

In `Pyro`, approximate inference is based on Markov Chain Monte Carlo (MCMC) or variational inference techniques (8) (4). While MCMC sampling faces computational challenges in scaling, variational inference depends heavily on the choice of the variational family. Practically, a number of structural assumptions, optimization, and scalability tricks have been demonstrated in both cases that allow these classes of algorithms to be scaled to large datasets. Machine learning frameworks such as Pytorch and Tensorflow have supported probabilistic inference through their respective libraries offering the ability to create and manipulate probability distributions (6) (7), but lack direct support for general inference. In contrast, probabilistic programming languages such as `Pyro` and Pyprob (3) are built on top of Pytorch's distributions module and offer a host of general-purpose inference algorithms that we will utilise in our project.

5. Modeling Policy Interventions

When governments deal with diseases, they may take certain measures that result in limiting the exposure of the population to the disease. These measures may range from mild as in washing hands to stringent as in enforcing a total lockdown. We have seen multiple government policies emerge that lie along this spectrum, all with the common goal of limiting the rising infections to acceptable levels. The goal of epidemiological modeling, particularly the spread of pandemics is to be able to infer the actions necessary to limit their spread. Every policy to address this is designed to intervene on the rate of spread

through natural or artificial means, for a short or long term duration. This motivates us to follow the work in (1) and explore the impact of such interventions in practice.

We consider a partially observed population and want to understand the putative controls that could achieve a goal defined as "control the spread of the disease", "reduce the death rate", or other such equivalent outcomes laid out in (1). While some of the parameters that define disease spread are controllable, there are certain non-controllable parameters including properties of the disease that we will infer. From a computational standpoint, we demonstrate the capability of the universal probabilistic programming language, `Pyro` (5), to combine a system of equations that define a simulator and perform inference over the latent variables within the simulation to obtain a posterior distribution over the model parameters. In addition to an inference engine, `Pyro` offers the capability to intervene on the variables within this simulation in order to obtain potential outcomes of policy changes that can be expressed within the language as fixing the values of certain parameters. We utilize this capability to craft intuitive experiments for an exposition of how different COVID-19 models and policies might evolve with time while remaining largely consistent with existing data. We utilized `Pyro` to do so on two compartmental models, SIR and SEIR.

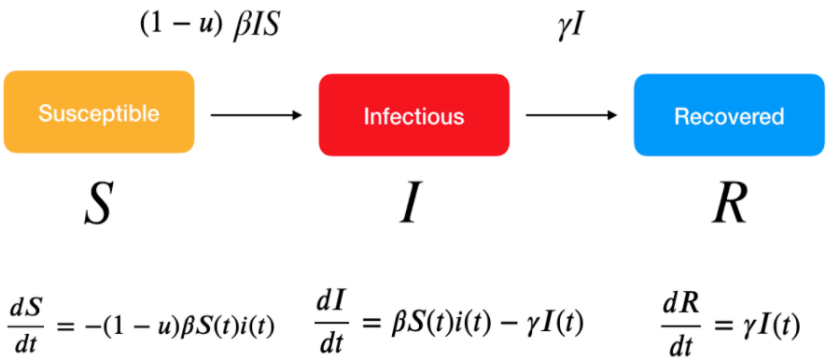


Figure 3: How policy intervention parameter u is included in compartmental model.

6. Modeling Noisy Observations

The authors define the population-level disease parameters we discussed for the SIR model (transmission probability β and recovery rate γ) in terms of certain measurable quantities such as the recovery time τ and the initial estimates of the reproduction number R_0 and response rate ρ. This allows us to place empirically informed priors on these latent variables and estimate their posterior distributions.

The response rate ρ indicates the proportion of observed infections in reality because we typically cannot expect to observe every single case of infection since people may not realize they have been infected unless there is extensive testing available, which has not been the case in most countries at least initially. The same may be exacerbated due to socioeconomic and political factors, so ρ allows us to model this under-reporting of COVID-19 cases in practice. Our ability to incorporate this into the model results in a far more realistic setting to fit using available external data (described in 8).

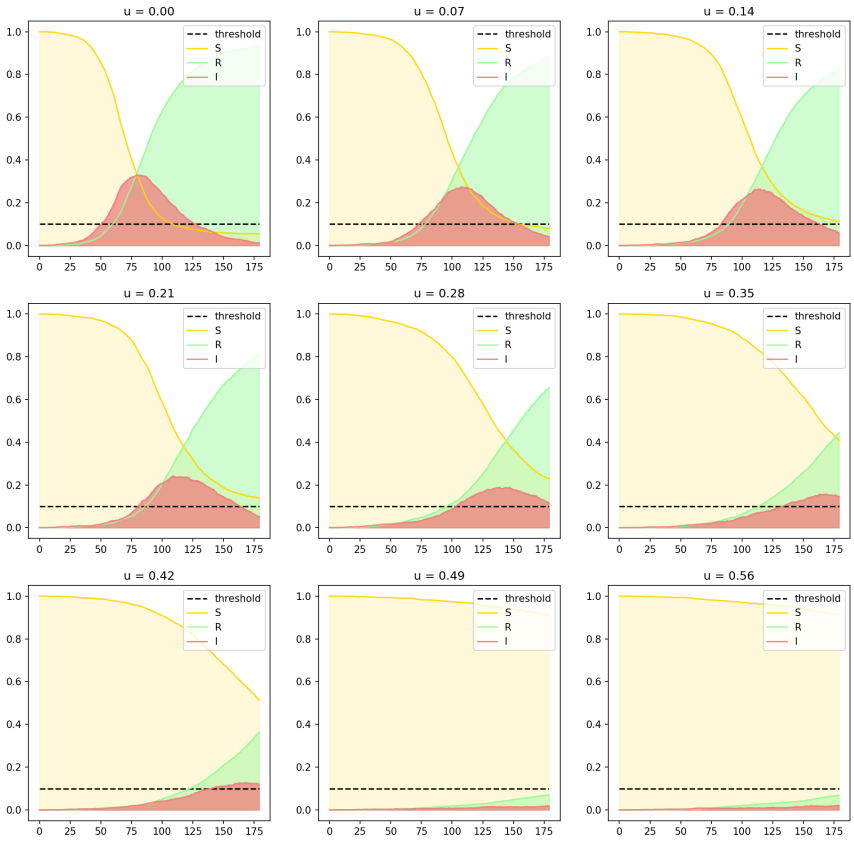


Figure 4: A grid visualizing the disease progression with time (along x-axis) under varying levels of the policy intervention parameter u for SIR.

7. Results

Our experiments yield posterior estimates 8 for the reproduction number $R0$ and the response rate rho as shown in the figure. We discover that the priors on $R0$ turned out to be quite reasonable since the posterior does not produce a huge change in their point estimates.

Finally, we successfully model disease progression and forecast for a set of future timesteps to recover potential new infections. We compare this to the actual values, and observe a strong correspondence between the two. We can re-introduce policy interventions at this stage and figure out the global minimally invasive intervention to perform that will remain within the desired thresholds for the infected populace.

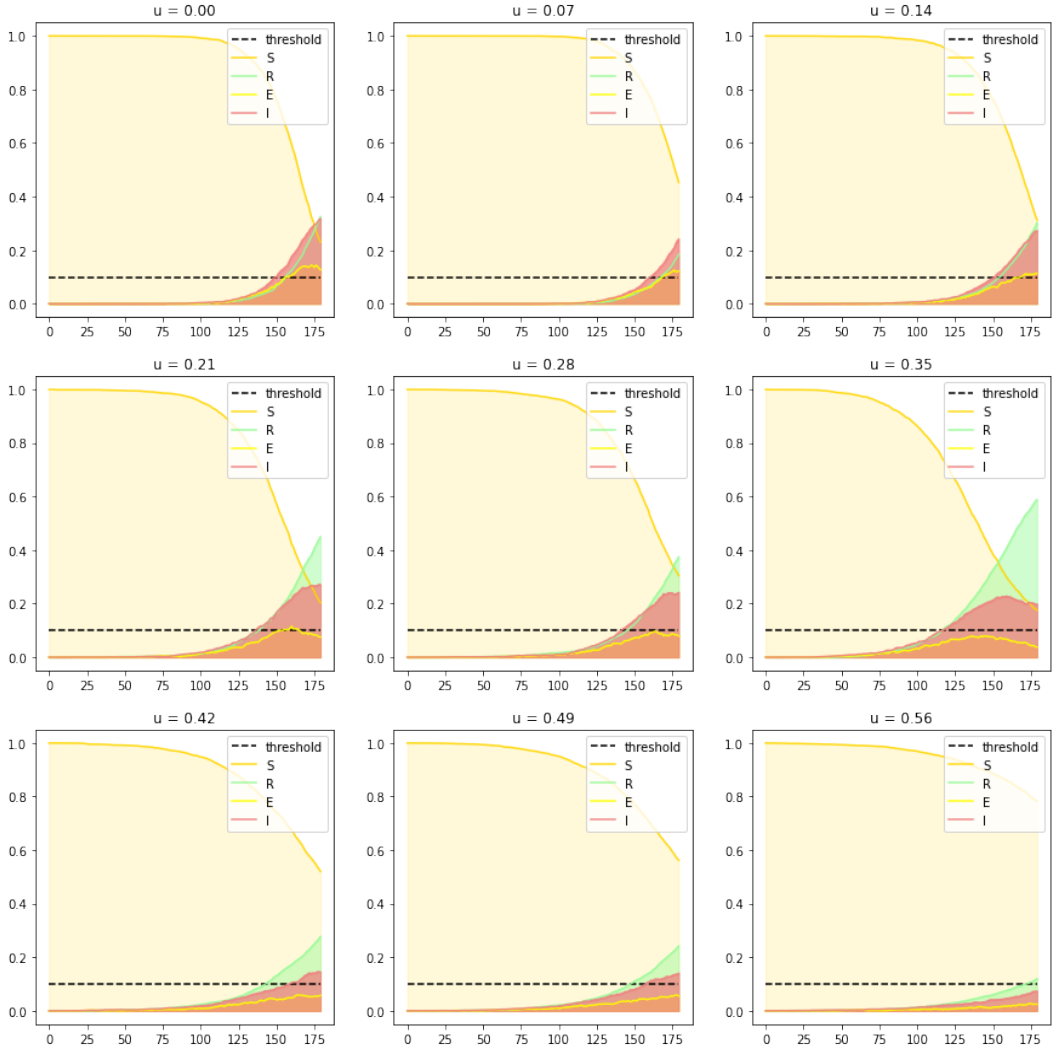
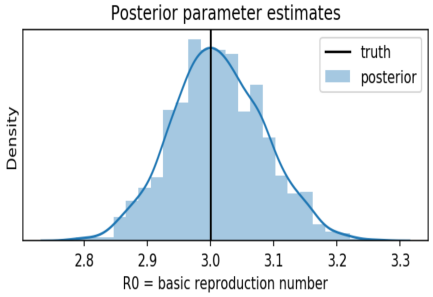


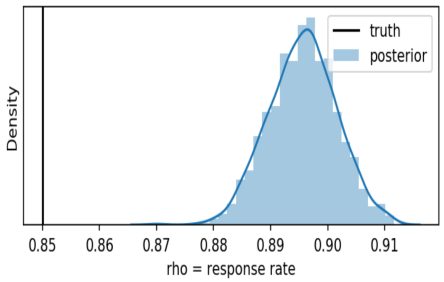
Figure 5: A grid visualizing the disease progression with time (along x-axis) under varying levels of the policy intervention parameter u for SEIR.

8. Ongoing Work

- In this report, we introduced policy interventions which represent the government implementing various policies to limit the spread of COVID-19. We considered a simple case where we want to know what is the minimally invasive constant policy corresponding to the value u_{min} that can be implemented to limit the spread of COVID right at the outset of the disease spread. We can easily implement more complex models for policy interventions that vary with time-step u_t such that the modified goal is to achieve a globally optimal sequence of policy interventions in order to model the infection spread. We are actively working on such a characterization of policy interventions.



(a) Posterior distribution for $R0$



(b) Posterior distribution for ρ

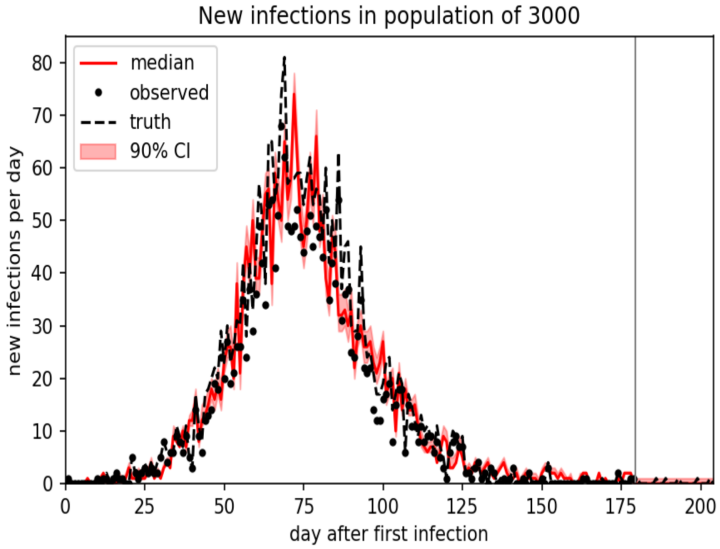


Figure 7: Predicting Future Infections (in this case, quite low)

- The work we have done thus far relies on creating a simulated source of data that belongs to the class of compartmental models. We mentioned agent-based simulators earlier in the context of having more granular interactions. The idea behind this is yet again to make weaker assumptions and remain closer to realistic scenarios. Agent-based simulations, unlike compartmental models allow us to vary individual-level features which results in a different set of dynamics in terms of disease spread. We have managed to successfully run the FRED influenza simulator utilised in (1) however omit the experiments since they are still incomplete. In general, we want to update the parameters of FRED to reflect COVID-like disease properties and obtain simulated data for fitting our probabilistic COVID-19 model. This yields better parameter estimates that may be closer to their real-world values.

- We are using real world data to fit the models but it is still a set of incomplete experiments which is why we leave it out of the report. The code and data to replicate

our experiments is however, available at the repository with the resources for this paper.

- We are recreating the SEI3RD model from (1) in Pyro. Their original diagram is provided **??** for visual understanding of the compartmental model. Correspondingly, we will have transitions defined between pairs of compartments that ultimately end up being closer to reality than the simpler models. We will augment this model by introducing noisy observations to accurately represent underreported cases of COVID-19.

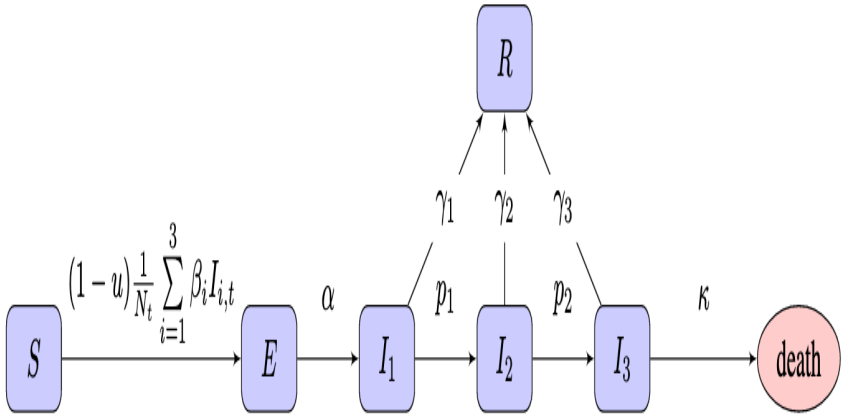


Figure 8: The SEI3RD Model we are recreating with policy interventions and noisy observations

References

[1] Wood, Frank, Andrew Warrington, Saeid Naderiparizi, Christian Weilbach, Vaden Masrani, William Harvey, Adam Scibior, Boyan Beronov, and Ali Nasseri. *"Planning as Inference in Epidemiological Models."* arXiv preprint arXiv:2003.13221 (2020).

[2] de Witt, Christian Schroeder, Bradley Gram-Hansen, Nantas Nardelli, Andrew Gambardella, Rob Zinkov, Puneet Dokania, N. Siddharth et al. *"Simulation-Based Inference for Global Health Decisions."* arXiv preprint arXiv:2005.07062 (2020).

[3] Le, T.A., Baydin, A.G. and Wood, F. *Inference Compilation and Universal Probabilistic Programming.* arXiv preprint arXiv:1610.09900, 2016.

[4] Kingma, D. P., & Welling, M. *Auto-encoding variational bayes.* The International Conference on Learning Representations, 2014.

[5] Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, Singh R, Szerlip P, Horsfall P, Goodman ND. *Pyro: Deep universal probabilistic programming.* The Journal of Machine Learning Research, 2019.

[6] Dillon, Joshua V., Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous *Tensorflow distributions.* arXiv preprint arXiv:1711.10604, 2017.

[7] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019 *PyTorch: An imperative style, high-performance deep learning library.* In Advances in Neural Information Processing Systems (pp. 8024-8035), 2019.

[8] Hoffman MD, Blei DM, Wang C, Paisley J. *Stochastic Variational Inference.* The Journal of Machine Learning Research, 2013.

[9] Baydin AG, Shao L, Bhimji W, Heinrich L, Meadows L, Liu J, Munk A, Naderiparizi S, Gram-Hansen B, Louppe G, Ma M. et. al. *Etalumis: Bringing probabilistic programming to scientific simulators at scale.* In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2019.