Pandemic Response as Reinforcement Learning

Blake Elias
New England Complex Systems Institute
Cambridge, MA
blakee@necsi.edu

Alexander F. Siegenfeld
Department of Physics
Massachusetts Institute of Technology
Cambridge, MA
asiegenf@mit.edu

Yaneer Bar-Yam
New England Complex Systems Institute
Cambridge, MA
yaneer@necsi.edu

Abstract

While pandemics begin as natural processes which can be *modeled*, they endure as anthropogenic ones which must be *steered* towards desired outcomes. A key gap in deciding how to intervene in pandemics has been confusion around the trade-off between economics and human health. We provide a means for discussing both epidemiological and economic concerns in a single framework, and frame one possible response — region-wide lock-downs — as a sequential decision-making problem (how much to close or re-open a region's economy over time). We show that an optimal policy can be determined using standard dynamic programming methods, and demonstrate that no static, "steady-state" policy can achieve optimal regret — rather, a real-time, dynamic policy is required.

We find, even with very adverse assumptions, that an optimal response includes strong, early action to reduce the level of viral spread, and subsequent relaxation of those restrictions, with rapid response when new outbreaks occur.

1 Introduction

While pandemics begin as natural processes, they endure as anthropogenic ones. Viral dynamics and human behavior form a joint system, which must be steered and managed. While epidemiology and theoretical biology provide a mature literature on **modeling and predicting** pandemic outcomes, the past few months have exposed a lack of consensus on how to **act** in the face of a pandemic.

In this work, we take one step towards closing this gap, treating both the public health and economic impacts of lock-downs under a unified framework of **decision-making under uncertainty.** We explore methods for reinforcement learning and optimal control as a framework for guiding our choices.

We discuss this in the language of the reinforcement learning (RL) literature, where an agent acts on its environment, observes new states, and experiences rewards (or penalties). In this setting, the state observations consist of the number of newly infected individuals at every time step. Costs will arise as a function of the number infected; there will also be costs associated with interventions, such as lock-downs.

We will speak of an agent that gets to act on the system, performing interventions to regulate how many infections occur. Typically an RL agent gets to directly act on the environment. In this case the primary agent is a policy-maker (e.g. a mayor, governor, etc.), whose actions impact the populous/citizenry. In turn, the populous itself is also an actor in the context of an outbreak. We abstract away this distinction by **treating both the policy-maker and the populous as a single agent** — which we denote as "the community" — whose combined behavior determines the trajectory of the outbreak.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

The community's task is to **find an optimal, dynamic policy** which maximizes reward over time. Should we do a strict, short lock-down, or a looser, long-lasting one? Should we pursue elimination, or merely keep new infections "under control" (i.e., a constant number of daily new cases)?

While there have been many analyses on particular strategies one might follow [1], we present a framework for finding a true optimal policy, using methods from optimal control — specifically, dynamic programming. Such computational techniques can lead to superior decisions that are sometimes counterintuitive.

2 The Problem

2.1 Environment (Infection Dynamics)

Consider a region with population N (e.g. a city, state, or even a small country). We model the system state as (S_t, I_t), where:

- S_t is the number of susceptible individuals at time t,
- I_t is the number of non-contained, infectious individuals at time t.

We consider a discrete, stochastic version of standard SIR dynamics (see [2] for a thorough treatment of stochastic epidemic models). We assume that any interventions to limit the growth of I_t, are being applied uniformly across the region. Hence, the behavior of the region's infections can be described by a single reproductive number, R_0. The expected number of new infections due to community transmission at time t, is then $\frac{S_t}{N} I_t R_0$ (each infected person is expected to infect another R_0 people, while only a $\frac{S_t}{N}$ fraction of those people are susceptible to infection). Finally, let α be the "importation rate": the expected number of new cases per time step that are imported from outside the region of interest, that do not get stopped by travel restriction or quarantine. (We assume this rate is constant.) We can now model the infection dynamics as:

$$\mathbb{E}[I_{t+1}] = \frac{S_t}{N} I_t R_0 + \alpha, \tag{1}$$

$$I_{t+1} \sim \text{NB}\left(k, \frac{k}{k + \mathbb{E}[I_{t+1}]}\right) \tag{2}$$

$$S_{t+1} = S_t - I_{t+1}, \tag{3}$$

where NB indicates the negative binomial distribution with over-dispersion parameter k. We use the negative binomial distribution to approximate "super-spreader" behavior [3, 4, 5, 6], with k estimated between 0.10 and 0.17.

Each time step represents 4 days. We assume that:

1. individuals are only infectious during a single time step,
2. any individuals *infected* at time t, will become *infectious* at time $t+1$.

We assume a low number of infections relative to the population size, rendering the effects of immunity negligible. Indeed, the number of infections required for immunity to become significant is so large, and the resulting cost so great (see section 2.3), that immunity effects are not critical in determining the optimal policy. Equivalently, individuals' limited risk tolerance has the potential to prevent such wide-scale spread in the first place.

2.2 Action Space

The agent in this scenario is the "community", representing the collective decisions of a government policy-maker as well as the populous. While there is a large space of choices (e.g. closing schools, transit, businesses, etc.), the relevant effect is that of changing R_0 in equation 1, to some value $R_t \in (0, R_0]$. Hence, we model the action space simply as the policy-maker setting R_t directly. Such actions will come at a cost: the lower R_t desired, the higher the cost. We discuss this further in the next section.

2.3 Reward Function

We recognize two sources of cost (i.e. negative reward) to the agent:

1. $C_I(I_t)$: the cost of infections at a single time step,
2. $C_R(R_t)$: the cost of locking-down.

The cost at each time step is the sum of these two costs:

$$C(t) = C_I(I_t) + C_R(R_t) \tag{4}$$

Cost of Infections: For a given number of infectious individuals $I = I_t$ at a particular time-step, we model the cost of these infections as:

$$C_I(I) = \begin{cases} aI & \text{if } I < H \\ aH + b(I - H) & \text{if } I \geq H \end{cases}$$

where:

- $H > 0$ represents the number of infections required to **overwhelm the hospital system**,
- $a > 0$ is the expected **cost per infected patient** when the hospital system is within capacity (this accounts for several factors such as the cost per hospital visit, fatality rate, long-term health consequences after recovery, etc.)
- $b > a$ is the cost per infected patient **once the hospital system is overwhelmed.** b accounts for the increased fatality rate, and worsened long-term health consequences, when patients do not get access to adequate medical care. This also accounts for any psychological penalty that society may place on the poor optics of having an overwhelmed hospital system, beyond the economic costs of the health outcomes themselves.

Cost of Lock-Down: We model the cost of setting R_t as a power-law:

$$C_R(R) = A/R^B,$$

where $A, B > 0$ are constant parameters determining the scale and curvature of this function. We use a power-law relationship to capture the fact that the cost of stopping an outbreak can approach infinity if one wishes to bring R_t down to 0, while no cost is incurred by letting R_t remain at its default, R_0.

2.4 Sequential Decision-Making Problem

Given the environment dynamics, action space, and reward function described above, the agent (policy-maker) is now faced with an optimization problem. Namely, to select the optimal $\{R_t\}_{t=1}^{T}$ that minimizes total cost, over some (possibly infinite) time-horizon T:

$$\min_{\{R_t\}_{t=1}^{T}} \sum_t \gamma^t C(t). \tag{5}$$

Here, $\gamma < 1$ is a discount-factor accounting for the **belief that the burden of the pandemic may lessen over time**. This could be due to the discovery of a highly effective vaccine, other medical treatments that make the expected loss per case much less, or the adaptation of society in finding ways to reduce contact transmission while maintaining economic activity. All such changes can be factored into the values of a, b, A, B, and H. E.g., $\gamma = 0.99$ represents the belief that there will be an effective, widely-administered vaccine 9 months from now (resulting in no future infections)—or, that all scaling parameters in the cost function (cost per case, i.e. a and b, and cost of locking down, i.e. A) will each get cut in half in that same time-frame.

3 Solution Method

We solve this problem using the value-iteration algorithm, as a method of solving Bellman's optimality equation [7]. We iteratively construct a value table, $V(S, I)$ representing the state-value of having

S currently susceptible and I currently infectious. Initializing $V^{(0)}(\cdot,\cdot)=0$, we iteratively apply Bellman's equation to convergence, yielding $V^*(\cdot,\cdot)$:

$$V^{(i+1)}(S_t,I_t)=\min_R\{\mathbb{E}[C(I_t,R)+\gamma V^{(i)}(S_{t+1},I_{t+1})]\}, \tag{6}$$

We then compute an optimal policy $R^*(\cdot,\cdot)$ with respect to this look-up table:

$$R^*(S_t,I_t)=\arg\min_R[C(I_t,R)+\gamma\mathbb{E}[V(S_{t+1},I_{t+1})]] \tag{7}$$

4 Experiments

We consider a range of possible values for α, A and B, investigate the optimal policy, and visualize response trajectories based on these policies. We use a population size of $N=1000$, $I_0=100$, $\alpha=0.5$, $A=70, B=1$, $a=b=1$, and $\gamma=0.99$. We set k, the dispersion parameter in the negative binomial distribution, to 10^{14} (approximating $k\to\infty$). Results are shown in the appendix.

We intentionally chose this set of parameters to make lock-downs quite expensive, and to make elimination quite difficult and un-favorable (described below). Nonetheless, we observe that the agent converges upon an elimination-seeking strategy. These parameters make elimination un-favorable for the following reasons:

- $\alpha = 0.5$ implies **one new imported case every two time steps** (i.e. every 8 days), in a population of 1000 people. At this rate, waiting 80 days would result in 1% of the population becoming infected from imported cases alone (even with no additional community transmission). This is a very high importation rate, making permanent elimination impossible to achieve. If the agent does want to pursue elimination, it will require recurring lock-downs.
- $A=70$ and $B=1$ make **reducing R to 1 just as costly as 7% of the population getting infected** in a single time-step. This makes the proposition of lock-down quite an expensive one.
- Letting $k\to\infty$ turns the negative binomial distribution of equation 1 into a Poisson — imposing **complete homogeneity of individual infectiousness**. The heterogeneity of infectiveness present in COVID-19, in fact makes elimination easier due to the possibility of die-out [8], while homogeneity makes elimination as hard as possible for a given R_0.
- The choice of $\gamma=0.99$ discounts all **costs 9 months from now as 50% cheaper than their cost today**. This could be due to better medical care, better testing, or the existence of a highly effective vaccine and its mass distribution. This is an optimistic assumption.

The factors above were designed to make elimination a difficult and unattractive option for the agent. Nevertheless, **the agent pursues an aggressive elimination strategy**. The nearest alternative strategy, of maintaining $R=1$ over a long time period (i.e. keeping daily new cases constant), accumulates higher cost over time due to the economic costs of continued lock-down and restrictions. Instead, by investing early-on to achieve near-elimination, the agent is later able to profit from re-opening the economy. The agent is willing to lock-down again as new outbreaks emerge — as many times as necessary, so long as new outbreaks continue to arise. We observed qualitatively similar results after performing sensitivity analysis on the various parameters, with $A\in\{100,500,1000\}, B\in\{1,1.5,2,2.5\}$, and $\alpha\in\{0.0,0.2,0.4,0.6,0.8\}$. Additional experiments, testing the impacts of immunity and infrequent interventions, will be forthcoming.

5 Conclusion

If we hope to make progress on the public health issue of pandemics, we will need clear ways to think about the interactions and trade-offs between human health and economics. To this end, we make the following contributions:

- We present **a unified framework** for considering both the economic and public health costs of a pandemic, and a normative framework for selecting optimal actions, in the frame of sequential decision-making and optimal control.
- We **instantiate several variants** of the optimal control problem, corresponding to the different biological and social realities we may face.

- We argue for choosing which problem variant and parameter regime to study, **based on worst-case assumptions.**
- We show, in one plausible parameter regime, that the **agent's optimal behavior is to pursue an aggressive elimination strategy** — in contrast to the variety of strategies we see most governments following currently — despite the high cost of achieving elimination.

Relevance to Workshop

This work is relevant to the workshop due to its use of ML methods (i.e. reinforcement learning) to guide the development of effective economic policy.

Prior Publication

The authors certify that this submission is original work that has not been published or first made available before January 1, 2017.

Broader Impact

The intended outcome of this work is to lessen the societal costs borne from pandemics, by bringing clarity to the dynamics of how different types of cost are linked, and presenting a framework for minimizing total cost. The policy choices surrounding such a situation have extraordinary ethical weight; this work provides one possible framework for thinking about such decisions in a principled way.

A positive outcome of this work would be to shape current and future public dialogue into a more cooperative mode, where all parties involved can agree on the criteria for assessing which strategy is best—an understanding that there exists a shared goal, and principled methods for optimizing towards that goal, such that all parties will be better off.

Potential negative outcomes could include an overly literal view that costs of equal value can always be exchanged. More broadly, there is some risk of over-committing to an economics-driven value system, which in some cases will not be the only value system to apply.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2020/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to autmoatically hide this section in the anonymized submission.

References

[1] S. R. Sheffield, A. York, N. A. Swartwood, A. Bilinski, A. Williamson, and M. C. Fitzpatrick, "Strict physical distancing may be more efficient: A mathematical argument for making lockdowns count," *medRxiv*, 2020.

[2] P. E. Greenwood and L. F. Gordillo, "Stochastic epidemic modeling," in *Mathematical and statistical estimation approaches in epidemiology*, pp. 31–52, Springer, 2009.

[3] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, "Superspreading and the effect of individual variation on disease emergence," *Nature*, vol. 438, no. 7066, pp. 355–359, 2005.

[4] A. Endo, S. Abbott, A. J. Kucharski, S. Funk, *et al.*, "Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china," *Wellcome Open Research*, vol. 5, no. 67, p. 67, 2020.

[5] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, *et al.*, "Early dynamics of transmission and control of covid-19: a mathematical modelling study," *The lancet infectious diseases*, 2020.

[6] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, *et al.*, "Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study," *The Lancet Infectious Diseases*, 2020.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[8] D. A. Kault, "Superspreaders help covid-19 elimination," *medRxiv*, 2020.

A Experimental Results

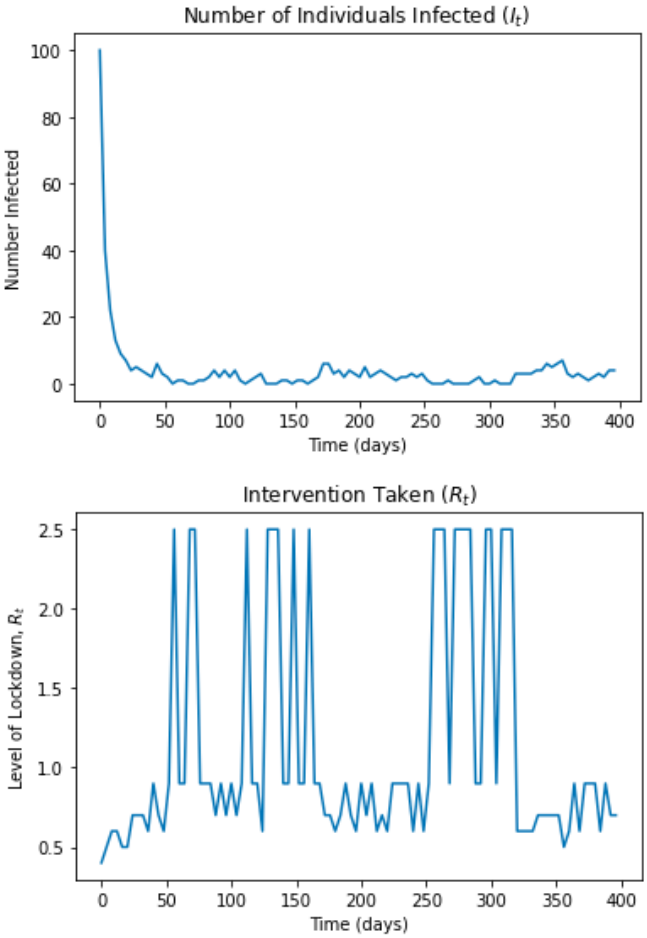

Figure 1: **Top:** trajectory of infections. **Bottom:** lock-down policy enacted over time. Agent pursues an aggressive elimination strategy with strict lock-downs followed by full re-openings.