Optimal, near-optimal, and robust epidemic control

Dylan H. Morris[1*], Fernando W. Rossine[1*], Joshua B. Plotkin[2], and Simon A. Levin[1]

[1]Department of Ecology & Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA
[2]Department of Biology & Department of Mathematics, The University of Pennsylvania, Philadelphia, PA, 19104, USA
[*]these authors contributed equally; correspondence to dhmorris@princeton.edu, frossine@princeton.edu.

April 3, 2020

Abstract

The COVID-19 pandemic has highlighted the need for control measures that reduce the epidemic peak ("flattening the curve"). Here we derive the optimal time-limited intervention for reducing peak epidemic prevalence in the standard Susceptible-Infectious-Recovered (SIR) model. We show that alternative, more practical interventions can perform nearly as well as the provably optimal strategy. However, none of these strategies are robust to implementation errors: mistiming the start of the intervention by even a single week can be enormously costly, for realistic epidemic parameters. Sustained control measures, though less efficient than optimal and near-optimal time-limited interventions, can be used in combination with time-limited strategies to mitigate the catastrophic risks of mistiming.

1 Introduction

Controlling a novel epidemic can be extremely difficult: there may be little standing immunity within the population, and the disease itself may be poorly understood. Such a disease can infect a large portion of the population and, if symptoms are severe, as in the ongoing COVID-19 pandemic, healthcare systems may be strained to the breaking point. It can take months to develop drugs, and years to develop vaccines [4]. Initial control efforts therefore rely heavily on "non-pharmaceutical interventions"[5] such as "physical distancing" (also called "social distancing") measures designed to reduce rates of disease-transmitting contact. These measures carry social and economic costs, and so policymakers may be unable to maintain them for more than a short period of time.

Here we derive the optimal strategy for limiting the peak prevalence of a novel disease using a time-limited intervention. We consider classes of easier-to-implement but less efficient strategies, and show that they perform nearly as well as the globally optimal strategy if they are themselves optimized. Importantly, neither the optimal strategy nor any of these near-optimal strategies is robust to implementation error: small errors in timing of the intervention produce large increases in peak prevalence.

However, we show that layering a short, strong intervention on top of a sustained weak intervention mitigates the risks of mistiming the strong intervention. Our results show that policymakers should be wary of trying to finesse an explosive epidemic of a novel pathogen by attempting the optimal intervention without any other controls. To avoid disasters born of implementation error, a strong, early, and ideally sustained response is required.

2 Background

2.1 Goal: peak reduction

Non-pharmaceutical control of an infectious disease is socially and economically costly, and it may be difficult to maintain due to political pressure and public non-compliance. A policymaker might be tempted to optimize, intending to maximize the benefits from a time-limited, highly efficacious intervention. To assess the prospects of this approach, including the risks of mis-implementing such a strategy, we first specify the goal to be achieved and determine a provably optimal intervention given that goal. We then analyze the costs incurred by errors in implementation of the optimal strategy.

In this article, we study strategies aimed at epidemic mitigation rather than at epidemic eradication. This makes our work particularly relevant for the study of a novel, emerging pathogen. For such a pathogen, the effective reproduction number $\mathcal{R}_e$ (the average number of new cases produced by each infectious individual prior to recovery) is often much larger than one in the absence of control. To eradicate the pathogen, we must sustain $\mathcal{R}_e < 1$ (so that each currently infectious person tends to produce less than one subsequent infection) for sufficient time to eliminate all cases. That may not be feasible, particularly during a global pandemic when re-introductions of cases from other regions and other countries can occur after the disease has been locally eradicated.

Since this work focuses on interventions that are constrained to be short, limited in efficacy, or both, we will assume policymakers are aiming at mitigation: "flattening the curve" [1]. We assume that vaccination and other control strategies that render individuals immune without having been infected are not yet available to our policymaker. Instead, interventions are limited to transmission-reducing measures such as physical distancing [7, 3]: temporary reductions in the rates of contacts that might cause transmission.

Our principal criterion for the success of an intervention will be *minimizing the highest peak of the epidemic*. The peak is the largest single quantity of infectious individuals at any one time. This quantity is critical because it is the point at which and health services will be most strained. An overwhelmed system can dramatically increase case-fatality rates and lead to increased rates of complications from infection [7, 3].

All else equal, a policymaker would also like to minimize the total cases during the epidemic, but for an explosively spreading novel virus, this consideration is secondary to minimizing peak incidence. Avoiding collapse of the healthcare system is the most valuable goal. What is more, among non-eradicative interventions, those that reduce the highest epidemic peak (almost) necessarily reduce the total case count (or "final size") of the epidemic, though they may not do so as efficiently as interventions specifically targeted at final size reduction [2].

2.2 Prior work on time-limited interventions

There have been relatively few modeling studies on time-limited strategies for peak reduction. One known result establishes that time-limited peak reduction interventions should start earlier than final size reduction interventions, all else equal [2]. But the optimal strategy to reduce the peak not known, nor is the robustness of such a strategy to implementation error.

2.3 The importance of timing

Given perfect and complete information, what would be the optimal time-limited intervention for reducing the height of the epidemic peak?

The peak incidence occurs when the $\mathcal{R}_e = 1$, that is, when new infections exactly balance recoveries. This point can be hastened—and the number of infections at that point thereby reduced—if the number of susceptible individuals in the population is depleted.

This is why time-limited intensive interventions cannot begin too early if they are to minimize the peak. Locking down transmission too early and too aggressively slows the depletion of susceptibles. If controls are then relaxed too soon, there will be a large second epidemic peak [2].

This gives us some intuition for the importance of timing, and for the properties of the optimal intervention: it must trade off depletion of susceptible pool against reduction in the rate at which people become infectious. We formalize this by finding a provably optimal peak-reduction strategy for the standard susceptible-infectious-recovered (SIR) epidemic model [6]. We also analyze near-optimal, but more realistic, intervention strategies; and we assess the robustness of interventions to implementation error.

3 Model

3.1 The standard SIR

We consider the standard normalized SIR model, which describes the fractions of susceptible $S(t)$, infectious $I(t)$, and recovered $R(t)$ individuals in the population at time t, where $S(t)+I(t)+R(t) = 1$ and $1 \geq S(t), I(t), R(t) \geq 0$. Each infectious individual has potentially

disease-transmitting contacts at rate $\beta \geq 0$, and recovers at rate $\gamma > 0$. In a large, well-mixed population the fractions of these individuals individuals change according to the ordinary differential equations:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta SI - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I \end{aligned} \tag{1}$$

We note that such a model has a basic reproduction number $\mathcal{R}_0 = \frac{\beta}{\gamma}$ and an effective reproduction number $\mathcal{R}_e = \frac{\beta}{\gamma} S(t)$. We will parameterize epidemics in terms of their $\mathcal{R}_0$ and γ, which together determine β. We analyze interventions that seek to minimize the peak prevalence (maximum fraction infected at any point in time during the epidemic). We will call this quantity $I^{\max}$.

3.2 Interventions $b(t)$

We consider interventions that reduce the effective rate of disease-transmitting contacts β. This is the impact of physical distancing measures like those that have been imposed in many countries in an attempt to curb the growth of the ongoing COVID-19 pandemic. Many other non-pharmaceutical interventions—as well as some pharmaceutical ones, such as antivirals that reduce "shedding"—also work by reducing β.

We permit interventions that operate on β only for some limited duration τ. We impose this constraint in light of the political, social, and economic impediments to maintaining an aggressive intervention indefinitely.

We model an intervention by defining a transmission reduction function $b(t)$, such that:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -b(t) * \beta SI \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= b(t) * \beta SI - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I \end{aligned} \tag{2}$$

If the intervention is initiated at some time $t = t_i$, it must stop at time $t = t_i + \tau$. So necessarily $b(t) = 1$ if $t < t_i$ or $t > t_i + \tau$. During the intervention (i.e. when $t_i \leq t \leq t_i + \tau$), we permit $b(t)$ to assume any value between 0 and 1, inclusive. This restriction assumes that we cannot intervene to "raise" β and thus $\mathcal{R}_0$ above their values of in the absence of intervention, but also optimistically assumes that $b(t)$ can be adjusted instantaneously, and that $\mathcal{R}_0$ can be reduced all the way to zero, at least for a limited time. In formal mathematical terms (see section B.3 of the Appendix), we require only that the intervention function $b(t)$ output values in the interval $[0, 1]$ and that it be right-continuous except for finitely many points of discontinuity.

4 The optimal intervention

We pose the following optimization problem: given the epidemiological parameters β and γ and the finite duration τ, what is the optimal choice of $b(t)$ for minimizing $I^{\max}$? Such an optimal intervention is of interest for two main reasons. It provides a standard point of comparison for evaluating other possible interventions. And it will allow us to show that optimized, time-limited responses have inherent risks and shortcomings—even in the best case scenario.

We prove (see Theorem 1 in the Appendix; all references to Theorems, Corollaries, and Lemmas are to the Appendix) that for any $\mathcal{R}_0$, γ, and τ there is a unique, globally **optimal intervention** $b(t)$ that starts at an optimal time t_i^{opt} and is given by:

$$b_{\text{opt}}(t) = \begin{cases} \frac{\gamma}{\beta S}, & t \in [t_i, t_i + f\tau) \\ 0, & t \in [t_i + f\tau, t_i + \tau] \end{cases} \tag{3}$$

The optimal approach to reducing $I^{\max}$ is to **"maintain and then suppress"**. We first spend a fraction f of our intervention time τ in a "maintain" phase: we choose $b(t)$ so that $\mathcal{R}_e = 1$. This maintains the epidemic at a constant number of infectious individuals equal to $I(t_i^{\text{opt}})$, while susceptibles are depleted at a rate $\gamma I(t_i^{\text{opt}})$. We then spend the remaining fraction $1 - f$ in a "suppress" phase, setting $\mathcal{R}_e = 0$ so that infectious individuals are depleted at a rate γI.

The effectiveness of the intervention, when it should commence, and the balance between maintenance and suppression all depend on τ: the total duration of the intervention. The longer we can intervene (larger τ), the more we can reduce the peak (smaller $I^{\max}$), the earlier we should optimally act (smaller t_i^{opt}), and longer we should spend maintaining versus suppressing (larger f) (Fig. 1 A, G, H, Theorem 1, Lemma 7, and Corollary 3).

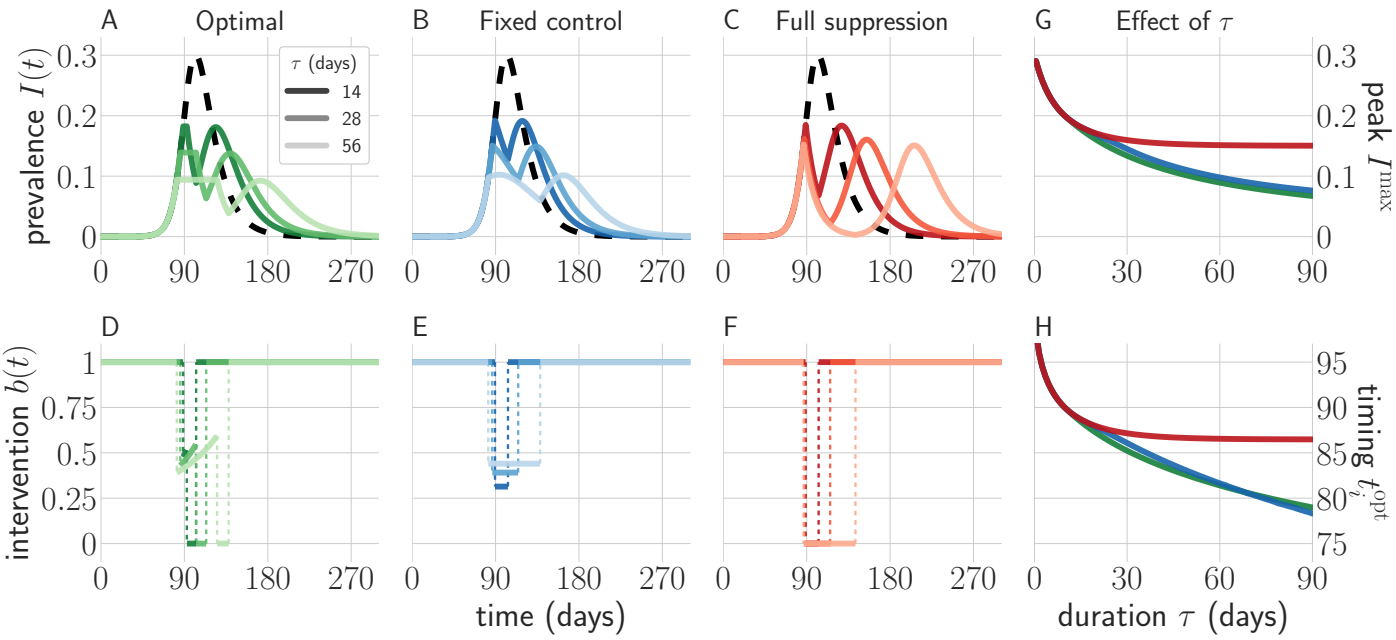

Figure 1: (A–F) Timecourses of epidemics under optimal (A), fixed control (B), and full suppression (C) interventions with three different values of τ: 14 days (dark lines), 28 days (intermediate lines), and 56 days (light lines), and their respective intervention functions (D–F). (G, H) Effect of the duration τ on the infectious prevalence peak (G) and intervention starting times (H) for the optimal intervention (green), the optimized fixed control intervention (blue), and the optimized full suppression intervention (red). Parameters as in Table 1.

5 Near-optimal interventions

Although optimal and theoretically enlightening, the intervention described above is unfeasible in practice. Its implementation would require policies flexible enough to fine-tune transmission rates continuously, imposing ever changing social behaviors. Implementation would also require instantaneous and perfect information about the current state of the epidemic in the population. We therefore also consider other families of potential interventions, and see how they perform compared to the optimal intervention.

Real-world interventions typically consist of simple rules that are fixed for some period of time (quarantines, restaurant closures, physical distancing). We model such **fixed control** strategies (previously investigated by di Lauro and colleagues [2]) as interventions of the form:

$$b_{fix}(t) = \sigma,\ t \in [t_i, t_i + \tau] \tag{4}$$

These fixed control interventions are fully determined by two parameters: the starting time t_i and the strictness $\sigma \in [0, 1]$. For any intervention duration τ we can then numerically optimize t_i and σ to minimize $I^{\max}$.

For a given $\mathcal{R}_0$, γ, and τ, an optimized fixed control intervention yields an epidemic time course that is remarkably similar to the one obtained under the globally optimal intervention strategy. Peak prevalence $I^{\max}$ is only slightly lower in the optimal intervention than in an optimized fixed intervention (Fig. 1 B G). Notably, the effectiveness of a fixed intervention depends on τ in a similar manner to that of the optimal intervention: longer interventions are more effective, should start earlier, and are less strict (Fig. 1 E, G, H).

The similarities between fixed control and optimal interventions can be understood by inspecting the time course of $\mathcal{R}_e$ during the intervention. A constant $\sigma > 0$ causes $\mathcal{R}_e$ to drop throughout the intervention. In this way, a fixed intervention initially depletes the susceptible fraction, but gradually starts depleting the infectious fraction as $\mathcal{R}_e$ falls. And so fixed control interventions are qualitatively similar to the "maintain-suppress" phases of the optimal intervention. As τ increases, the optimal σ decreases, promoting interventions that have longer "maintenance"-like periods characterized by susceptible depletion. That is, optimizing a fixed control strategy has the effect of choosing σ and t_i that emulate the optimal strategy. This emulation is in fact very successful, and leads an optimized fixed control intervention to perform nearly as well as the optimal intervention across a large range of durations τ (Fig. 1, G).

To further investigate the importance of the "maintenance" phases of these interventions, we consider a third class of interventions: the **full suppression interventions** defined by

$$b_0(t) = 0,\ t \in [t_i, t_i + \tau]. \tag{5}$$

These interventions are fully determined by the starting time t_i. They can be optimized for any intervention duration τ by choosing t_i appropriately. Such interventions emulate extremely strict quarantines. They are characterized by the complete absence of susceptible depletion during the intervention.

Note that the full suppression intervention is a limiting case both of the optimal intervention with vanishing short "maintenance" phase ($f \to 0$) and of the fixed control intervention with maximal strictness ($\sigma = 0$). Accordingly, the full suppression intervention performs similarly to the optimal intervention and to the fixed intervention for short durations, when those favor a relatively short "maintenance" phase and a high strictness (Fig. 1 C). For longer interventions, the effectiveness of full suppression rapidly plateaus at $I^{\max} = \frac{1}{2} + \frac{1}{2\mathcal{R}_0}\left(\log\left(\frac{1}{\mathcal{R}_0}\right) - 1\right)$ (Fig. 1 G, Corollary 5). Accordingly, the starting time of a full suppression intervention with increasing duration also plateaus: there is no benefit in fully suppressing too early (Fig. 1 H, Corollary 5).

Taken together, these results show that the most efficient way for long interventions to decrease peak prevalence $I^{\max}$ is to cause susceptible depletion while limiting how much the number of infectious individuals can grow. For short interventions, it is most efficient simply to reduce the number of infectious individuals.

Optimizing an intervention trades off cases now against cases later. We prove (Theorem 2) that maintain-suppress interventions optimized given f (which includes both the globally

optimal intervention and the full suppression interventions) cause the epidemic to achieve the peak prevalence exactly twice: once during the intervention and once immediately afterward. We conjecture that this result holds for optimized fixed interventions as well.

6 Mistimed interventions

The optimized interventions we have described are extremely powerful. For the COVID-like parameters shown, 28-day optimal or fixed control intervention can reduce the peak prevalence from about 30% of the population to under 15%. Even the full suppression intervention reduces peak prevalence to well under 20%. These are massive and potentially health system-saving reductions.

But interventions are not automatically triggered at a certain number of infectious individuals or a certain point in time. They are made by policymakers, who must estimate the current quantity of infectious individuals $I(t)$, often from very limited data, must begin roll-out of an intervention without certainty about how long that roll-out will take, and must also estimate the epidemiological parameters $\mathcal{R}_0$ and γ. These tasks are difficult, and so policymakers may fail to intervene precisely at the optimal moment t_i^{opt}.

Indeed, the COVID-19 pandemic has highlighted how difficult real-time epidemiological inference and response can be. Large numbers of asymptomatic and mildly symptomatic cases [8], as well as difficulties with testing, particularly in the United States [10], have left the public health community with substantial uncertainty both regarding the virus's epidemiological parameters and regarding case numbers in many locations.

All of this means that even if we grant our policymaker the capacity to tune $b(t)$ instantaneously and with infinite precision so as to maintain the optimal intervention, the policymaker will still intervene with some timing error. The realized t_i will be either greater or smaller than t_i^{opt}. To study this, we will consider errors in both directions, though there are reasons to believe that lateness might be more common than earliness, due to political and regulatory difficulties in intervening, or unexpected delays in implementation once the go signal is given.

How costly is mistiming a time-limited optimized intervention?

We find that even a single week of separation between the time of intervention and t_i^{opt} can be enormously costly, for realistic epidemic parameters (Fig. 2).

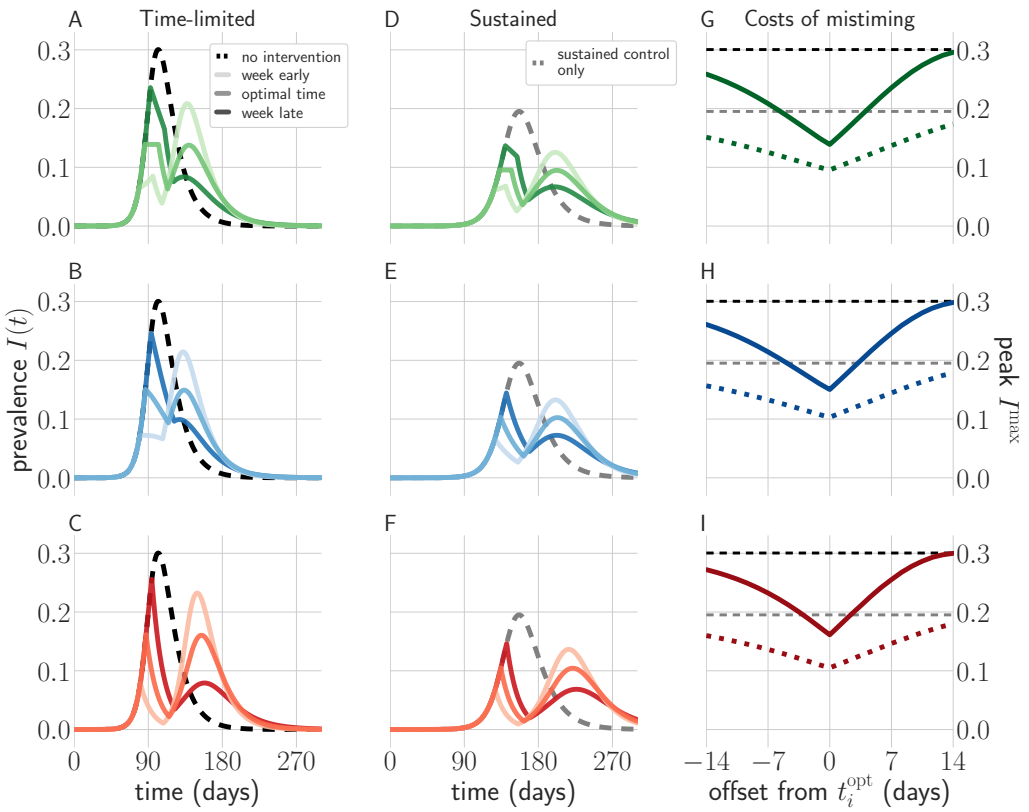

Figure 2: (A–F) Timecourses of epidemics under optimal (A), fixed control (B), and full suppression (C) interventions that are possibly mistimed: a week late (dark lines), optimally timed (intermediate lines), and a week early (light lines). Dashed black line shows timecourse in the absence of intervention. (D–F) Timescourses of epidemics with a sustained control that reduces $\mathcal{R}_0$ by 25%, combined witht the effects of a possibly mistimed optimal (D), fixed control (B), and full suppression (F) interventions, with line lightenss as before. Dashed grey line shows timecourse with only sustained control and no additional intervention. (G–I) Effect of offset of intervention time t_i from optimal intervention time t_i^{opt} on epidemic peak prevalence $I^{\max}$ without (solid lines) and with (dotted lines) sustained control for optimal (G), fixed control (H) and full suppression (I) interventions. Dashed black and grey lines show $I^{\max}$ in the absence of intervention, without and with sustained control, respectively. Unless otherwise stated, parameters as in Table 1.

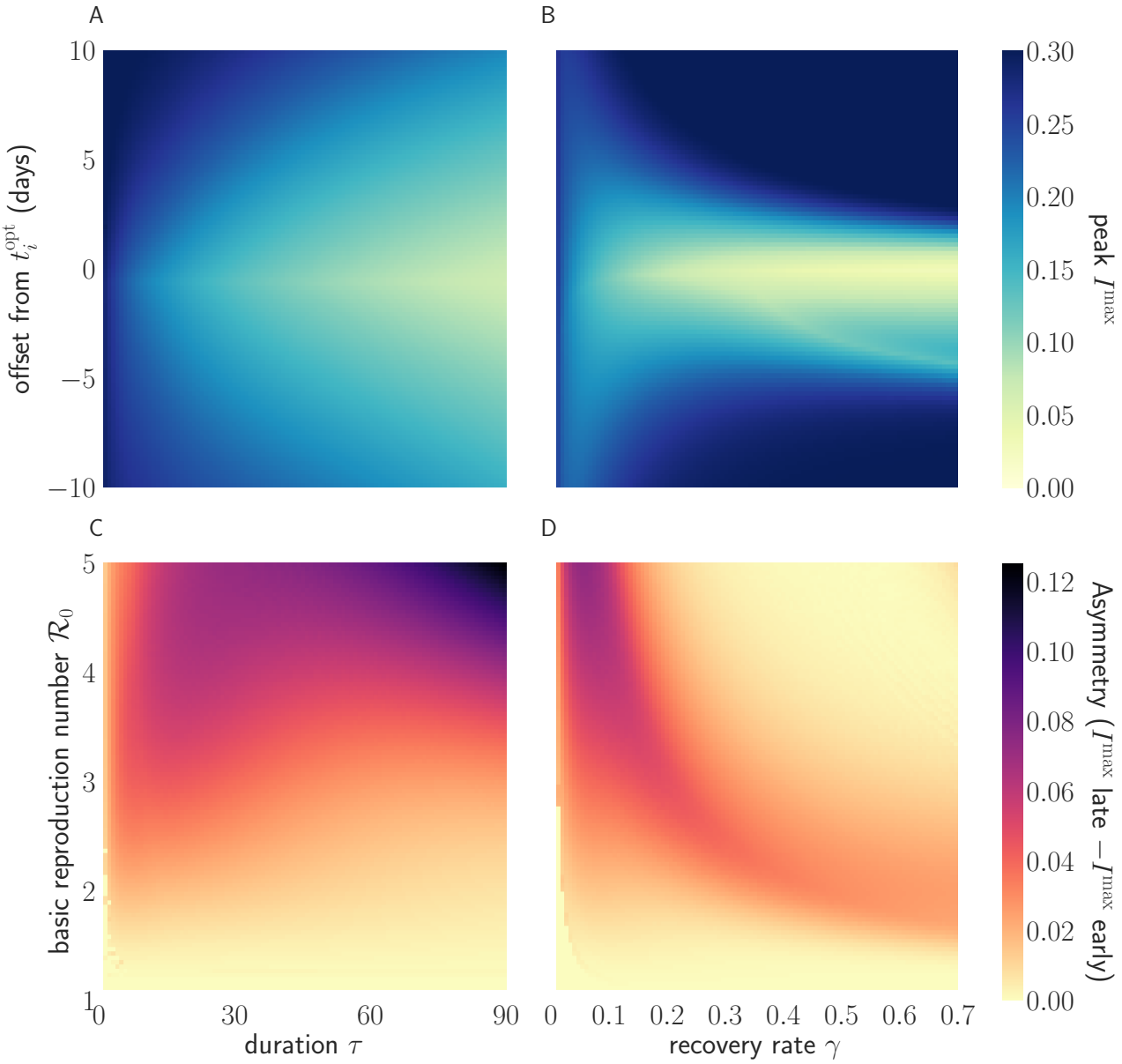

Figure 3: Effect of parameter variation on peak reduction and robustness to mistiming. (A, B) peak $I^{\max}$ as a function of offset from t_i^{opt} and (A) duration τ, (B) recovery rate γ. (C,D) Asymmetry between being intervening early and intervening late, quantified as $I^{\max}$ for a 7-day late intervention minus $I^{\max}$ for a 7-day early intervention, as a function of the basic reproduction number $\mathcal{R}_0$ and (C) duration τ, (D) recovery rate γ. Parameters as in table 1 unless otherwise stated.

While the optimal intervention achieves a dramatic reduction in the height of the peak,

mistiming the intervention can be disastrous. Intervening too early produces a resurgent peak, but it is even worse to intervene too late. For example, if the intervention is initiated one week later than the optimal time, then $I^{\max}$ is barely reduced compared to the absence of any intervention whatsoever, particularly in the case of a full suppression intervention.

The extreme costs of mistiming arise from the steepness of the $I(t)$ curve at t_i^{opt}. Optimized interventions permit some cases now in order to reduce cases later. Both infectious depletion and susceptible depletion require the presence of currently infectious individuals in order to be effective peak reduction mechanisms. This means that, except for interventions of very long duration τ, the optimal start time t_i^{opt} is during a period of rapid, near-exponential growth in the fraction infectious $I(t)$. Indeed, for epidemics that have faster dynamics, the consequences of mistiming interventions are increasingly stark (Fig. 3, B).

The problem with this is intuitive: because $S(t)$ and $I(t)$ are so steep at $S(t_i^{\text{opt}}), I(t_i^{\text{opt}})$, small errors in timing produce large errors in terms of $S(t_i), I(t_i)$. But the error is also asymmetric: being late is costlier than being early (Fig. 3 B–D). The early intervention is strongly sub-optimal and allows a large resurgent second peak, but that resurgence is still slower and smaller (thanks to susceptible depletion during the too-early intervention) than the prolonged uncontrolled growth period that is permitted by a late intervention (Fig. 2, A–C, G–I, Fig. 3).

Importantly, Theorem 2 implies that the late implementation of an optimized intervention leads to an immediate epidemic peak, whereas the early implementation of such an intervention postpones the peak (Fig 2, A-C). In practice this means that early interventions allow for course correcting. Interestingly, the optimal intervention displays a partial self correction for early implementation even—in fact, especially—for very fast epidemics. This occurs because during the initial phase of the optimal intervention, the strictness actually decreases in time (see equation 12), which allows for some growth of the infectious fraction and improved depletion of the susceptible fraction. An intriguing side effect of this automatic course-correction is that for relatively fast epidemics, some premature interventions outperform others that are less early (note the branching in Fig 3, B).

7 Sustained interventions

Are there any sustained measures that can improve the robustness with respect to timing of the optimized interventions? Because the severity of mistiming is governed by the steepness of $I(t)$, measures that reduce such steepness should alleviate the impact of mistiming an intervention. We propose using less strict measures of a longer duration to achieve this. Even though these measures by themselves might not strongly affect $I^{\max}$, they might buffer timing mistakes when used in combination with stricter, time-limited interventions.

We consider both a sustained control intervention, modeled as a constant reduction of $\mathcal{R}_0$ throughout the entire epidemic, as well as a time-limited intervention layered on top, during the same epidemic (Fig. 2, D–F). If perfectly timed, the limited interventions by themselves outperform the sustained intervention. Moreover, the addition of a sustained intervention to a perfectly timed limited intervention provides little extra benefit in terms of reduction of $I^{\max}$ (Fig. 2 G–H). However, even a slight mistiming of time-limited interventions makes them worse than a sustained intervention on its own. Most importantly, if both sustained

and time-limited interventions are adopted, the time sensitivity of the time-limited interventions is reduced. This is particularly pertinent for interventions that are later than optimal (Fig. 2 D–I).

8 Discussion

The SIR model is nearly a century old [6], and to this day it is used to inform policy, including the ongoing response to COVID-19. COVID-19 has also thrown into relief the importance not merely of minimizing total cases but of "flattening the curve". Our work establishes a provably optimal approach to curve-flattening in an SIR, given a limited intervention duration. We also show that coarser time-limited interventions can closely approximate the optimal intervention, suggesting a tantalizing policy opportunity. For a policymaker looking to minimize social and economic disruption during disease control, swallowing the bitter pill of an intensive intervention but otherwise being able to maintain business as usual might seem extremely attractive.

But we have also seen the costs that optimization can bring in practice, when timing errors will surely occur. Missing the optimal moment t_i^{opt} leads to a dramatic peak incidence. It is particularly costly to be too late; a high peak is achieved rapidly and immediately, before the health system has time to prepare.

There are many reasons a policymaker could miss t_i^{opt} even in the (already unrealistic) case that the underlying epidemiological parameters are perfectly known. For example, one must estimate the current number of cases in order to know the effective value of t itself, and thus how far away t_i^{opt} is. One must also introduce and enforce control measures—full compliance could easily take longer to achieve than planned.

For respiratory viruses, another concern is that $\mathcal{R}_0$ is not truly fixed in time even in the absence of intervention. $\mathcal{R}_0$ for a respiratory virus tends to exhibit "seasonality": rising in temperate zone winters, falling in temperate zone summers. Influenza is the classic example: the Northern Hemisphere has a winter "flu season". Poor estimates of seasonal variation could also disrupt attempts to optimize $b(t)$.

In ongoing work, we aim to explicitly analyze this problem of disease control under uncertainty, by coupling our intervention approaches with an analysis of epidemiological inference. If one must determine and act at t_i^{opt} based on noisy epidemiological data, how well can one do, and how are errors of inference magnified by the subsequent intervention mistiming?

We have assumed that $b(t)$ can be chosen at will, provided $0 \leq b(t) \leq 1$. But in practice $b(t)$ can be tuned at best coarsely; even the fixed interventions are not truly enforceable in their idealized form. Moreover, in practice more severe interventions will carry greater costs. An explicitly control-theoretic approach to the problem that trades off the costs of severe time-averaged interventions against their peak-reducing benefits would also be of substantial interest, and could also benefit from an analysis of robustness to implementation error.

Our work carries with it a number of other caveats, and leaves a great body of other important questions for additional investigations. While the epidemic peak is a good proxy for demands on healthcare system capacity, an even better metric would be the total person-days over healthcare capacity: the area A between the $I(t)$ curve and the line (or curve) of

maximal system capacity c. Any intervention that brings $I^{\max} < c$ necessarily sets that area to 0, but when choosing among strategies that cannot reduce $I^{\max}$ below c, the strategy that minimizes $I^{\max}$ will not necessarily minimize A. It might be preferable to permit a slightly higher peak but then take a full suppression approach, leading to a spike with small area above healthcare capacity. In general, however, we observe that as c approaches $I^{\max}$, the problem of minimizing the area A becomes almost identical to the problem of minimizing $I^{\max}$. When c is far from $I^{\max}$, it is closer to the problem of minimizing the final size of the outbreak, and strategies designed for final size reduction [2] may be superior to strategies designed for peak reduction.

We have also seen (Fig. 2) that despite their suboptimality from the point of view of reducing $I^{\max}$, premature interventions delay substantially the time until $I^{\max}$ is achieved. Peak delay may in some cases be a more pressing aim than peak reduction, if, for instance, healthcare capacity can be increased in the interim. In those situations, our results suggest that an early response weighted more toward suppression than maintenance would likely be desired, but strategies targeted specifically at peak delay should be analyzed in their own right. We also note that peak delay is another benefit of sustained control, even when that sustained control is relatively weak.

That said, our results, though simple, offer several robust and practicable principles for policymakers:

8.1 Principle 1: act early

There is an asymmetry between reacting "too early" and reacting too late. The optimal intervention time t_i^{opt} is poised at a cliff's edge. As γ increases that cliff only steepens; there is an sharper transition between doing optimally and doing terribly (Fig. 3 B). The costs of being early increase more slowly with the degree of error. Moreover, while we do not model this, early action leaves time for a course correction if it is too strong, and delays the higher second peak that it permits. Finally, our analysis shows that early intervention is to a degree self-correcting: the higher-than-expected stock of susceptibles when the intervention begins permits the epidemic to grow more or less rapidly up to the intended $I(t_i)$.

8.2 Principle 2: slow things down

It is easier to time a less steep exponential. Implementing moderate physical distancing measures can slow case growth, which allows for robustly timed aggressive interventions when they are needed. We also suspect that slowing the growth curve will make the inference of epidemiological parameters easier and more accurate, improving one's chance of hitting t_i^{opt} even as it reduces the costs of missing it. We will study this in future work.

8.3 Principle 3: when all seems lost, bear down

The remarkable success of the crude full suppression interventions in reducing peaks and delaying what they cannot reduce suggests that a policymaker who has evidence that they may be acting late should bear down as soon as possible with a policy as close to full

suppression as is achievable. For a disease like COVID-19, in which today's case data is in fact a snapshot of the situation one to two weeks earlier, what feels right on time may in fact be too late, and what feels late may be disastrously late. With that in mind, a policymaker looking to salvage matters should take a full suppression approach. In fact, full suppression is good as any other strategy at minimizing $I^{\max}$ whenever $t_i > t_i^0 > t_i^{\text{opt}}$, where t_i^0 is the optimal t_i for a full suppression intervention (Corollary 6), since $I^{\max}$ will simply be $I(t_i)$. Moreover, if we are seeking to minimize total person-days above capacity, full suppression will be first among equals, as it is by definition the fastest way to reduce $I(t)$ and get down off the peak.

9 Conclusion

Optimized time-limited interventions are extremely effective and efficient ways to reduce the peak of an epidemic. Deriving the form of optimal strategies is illuminating. It highlights the fundamental materials—susceptible depletion and infectious depletion—of any epidemic mitigation strategy. But these interventions are not robust to implementation error, and that is reason to rethink reliance on them alone, particularly when limited to incomplete information and combating a novel, poorly understood disease. Real-world policy must emphasize not efficiency but robustness.

10 Code availability

All code needed to reproduce numerical results and figures is archived on Github (https://github.com/dylanhmorris/optimal-sir-intervention) and on OSF (https://osf.io/rq5ct/), and licensed for reuse, with appropriate attribution/citation, under a BSD 3-Clause Revised License.

11 Acknowledgements

We thank Ada W. Yan, Amandine Gamble, Corina E. Tarnita, Elizabeth N. Blackmore, James O. Lloyd-Smith, and Judith Miller for helpful comments on previous versions of this work. We thank Juan Bonachela for helpful discussions.

DHM and SAL gratefully acknowledge financial support from NSF grant CCF 1917819.

12 Competing interests

We have no competing interests to declare.

A Further methods

A.1 Parameter choices

Table 1: Model parameters, default values, and sources/justifications

Parameter	Meaning	Units	Value	Source or justification
$\mathcal{R}_0$	basic reproduction number	unitless	3	Estimates for COVID range from 2 to 3.5 [8, 9]
γ	recovery rate	1/days	$\frac{1}{14}$	Infectious period for COVID of approximately 1–2 weeks [12]
β	disease-causing contact rate	1/days	$\mathcal{R}_0\gamma$	calculated
τ	duration of a time-limited intervention	days	28	Approximately a month

A.2 State-tuned and time-tuned maintain-suppress interventions

When we ask what it means for a maintain-contain style intervention to be mistimed, we need a model of how the intervention is implemented. One possibility is that the policymaker directly observes $S(t)$ throughout the intervention and chooses $b(t) = \frac{\gamma}{\beta S(t)}$ based on the directly observed $S(t)$. We call this a **state-tuned intervention**. Alternatively, the policymaker *plans* to intervene at some value S_i predicted to occur at t_i and chooses the values $b(t)$ knowing that if the intervention does indeed begin at $S_i = S(t_i)$, $S(t)$ will equal $S_i - \gamma I_i(t - t_i)$ during the maintenance phase. The policymaker then chooses $b(t)$ according to that *predicted* $S(t)$. We call this a **time-tuned intervention**. When we study mistimed interventions in the main text, we use time-tuned intervention; we see time-tuned interventions as a more realistic model of how a maintain-suppress intervention, if possible at all, would in fact be implemented, since instantaneous epidemiological observation is not possible. If it were possible, it could in fact permit a timing error to be better mitigated than state-tuning itself allows—by choosing the optimal strategy conditional on intervening at the true (S_i, I_i). Indeed, it can be seen that time-tuned interventions are in fact slightly more robust to mistiming than state-tuned interventions, as they are partially self-correcting where the state-tuned interventions are not (see main text section 6).

B Theorems and proofs

B.1 Useful notation

We define S_{crit} for an SIR model to be the critical fraction susceptible at which $\mathcal{R}_e = 1$ and $\frac{\mathrm{d}I}{\mathrm{d}t} = 0$ (in the absence of intervention), i.e. $S_{\text{crit}} \equiv \frac{1}{\mathcal{R}_0}$.

B.2 Maximum value of an SIR

Define $I^{\max}(t)$ for an SIR system as the maximum value of $I(x)$ achieved on the interval $x \in [t, \infty)$. Notice that $I^{\max}(0) = I^{\max}$, where $I^{\max}$ is the global maximum value of $I(x)$, which we are seeking to minimize with our intervention.

A known result that is immediate from the original work of Kermack and McKendrick [6, pp. 712-715] (see [11] for an explicit derivation) holds that for $S(t) \geq S_{\text{crit}}$:

$$I^{\max}(t) = I(t) + S(t) - \frac{1}{\mathcal{R}_0} \log\Big(S(t)\Big) - \frac{1}{\mathcal{R}_0} + \frac{1}{\mathcal{R}_0} \log\Big(\frac{1}{\mathcal{R}_0}\Big) \tag{6}$$

Remark 1: Continuity of $I^{\max}$. *I^{max} is a continuous function $I^{max} : [0,1]^2 \mapsto [0,1]$ of $v = (S(t), I(t))$ for $v \in [S_{crit}, 1] \times [0,1]$, and therefore also a continuous function of t, $I^{max} : \mathbb{R} \mapsto [0,1]$ for $t \in (-\infty, t_{crit}]$, where t_{crit} is the time such that $S(t_{crit}) = S_{crit}$*

This is immediate from the fact that $I^{\max}(t)$ is a linear combination of univariate functions of $S(t)$ and $I(t)$ that are themselves continuous on $[S_{\text{crit}}, 1]$ and $[0, 1]$, respectively. And since $S(t)$ and $I(t)$ are continuous functions of t, $I^{\max}$ is a continuous function of t for $t \in (-\infty, t_{\text{crit}}]$.

Lemma 1: The more fire, the bigger the blaze. *If $0 \leq I_x(t_x) \leq I_y(t_y)$ and $S_x(t_x) = S_y(t_y)$ for two SIR systems x and y with identical parameters $\mathcal{R}_0$ and γ at possibly distinct times $t_x, t_y \in [0, \infty]$ then $I_x^{max}(t_x) \leq I_y^{max}(t_y)$, with equality only if $I_x(t_x) = I_y(t_y)$.*

Proof. There are three cases.

Case 1: $I_x(t_x) = 0$. In this case, $I_x^{\max}(t_x) = 0 \leq I_y(t_y) \leq I_y^{\max}(t_y)$, with equality only if $I_y^{\max}(t_y) = 0$, which can only occur if $I_y(t_y) = 0$.

Case 2: $S_x(t_x) = S_y(t_y) < S_{\text{crit}}$. In this case, $I_x^{\max}(t_x) = I_x(t_x)$ and $I_y^{\max}(t_y) = I_y(t_y)$, so our result holds.

Case 3: $S_x(t_x) = S_y(t_y) > S_{\text{crit}}$. In this case, we can apply equation 6. Fixing $S(t) > S_{\text{crit}}$, $I^{\max}(t)$ is an increasing function of $I(t)$ (there is a single, positive $I(t)$ term in the sum), so the result must hold. □

Lemma 2: The more fuel, the bigger the blaze. *If $I_x(t_x) = I_y(t_y) > 0$ and $S_x(t_y) \leq S_y(t_y)$ for two SIR systems x and y with identical parameters $\mathcal{R}_0$ and γ at possibly distinct times $t_x, t_y \in [0, \infty]$ then $I_{t_x}^{max} \leq I_{t_y}^{max}$, with equality only if $S_x(t_x) = S_y(t_y)$ or $S_y(t_y) < S_{crit}$*

Proof. There are three cases.

Case 1: $S_y(t_y) \leq S_{\text{crit}}$. If $S_y(t_y) \leq S_{\text{crit}}$, then $S_x(t_x) \leq S_{\text{crit}}$, and neither epidemic will grow after t_x or t_y, respectively. It follows that $I_{t_x}^{\max} = I_x(t_y) = I_{t_y}^{\max} = I_y(t_y)$

Case 2: $S_y(t_y) > S_{\text{crit}}$, $S_x(t_x) < S_{\text{crit}}$. In this case, $I^{\max}_{t_x} = I_x(t_y)$ but $I^{\max}_{t_y} > I_y(t_y)$, since $\frac{\mathrm{d}I_y}{\mathrm{d}t} > 0$ if $I_y(t) > 0$ and $S_y(t) > S_{\text{crit}}$. So we have $I^{\max}_{t_x} < I^{\max}_{t_y}$

Case 3: $S_x(t_x), S_y(t_y) > S_{\text{crit}}$. The result in this case follows immediately from the fact that, fixing $I(t)$, $I^{\max}(t)$ is an increasing function of $S(t)$ when $S(t) > S_{\text{crit}}$. We can see that it is by taking the partial derivative the expression from equation 6 with respect to S:

$$\frac{\partial}{\partial S} I^{\max}(t) = 1 - \frac{1}{\mathcal{R}_0 S} \tag{7}$$

If $S > S_{\text{crit}}$, $\frac{1}{\mathcal{R}_0 S} < 1$, so $\frac{\partial}{\partial S} I^{\max}(t) > 0$, and $I^{\max}(t)$ is an increasing function of $S(t)$. □

B.3 Intervention function

We say that a right-continuous function with finite discontinuity points $b(t) : \mathbb{R} \mapsto [0, 1]$ is an intervention beginning at t_i with duration τ if $b(t) \equiv 1$ for all $t \in (-\infty, t_i) \bigcup (t_i + \tau, \infty)$. The SIR model under such an intervention will then take the form

$$\begin{aligned} \frac{dS}{dt} &= -b(t) * \beta SI \\ \frac{dI}{dt} &= b(t) * \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{8}$$

We wish to show that for every τ there exists an intervention that minimizes $I^{\max}$ and that such an optimal intervention must be identical except at a finite set of times $\{t_j \mid t_j \in (t_i, t_i + \tau)\}$ to one of the form:

$$b_{\text{opt}}(t) = \begin{cases} \frac{\gamma}{\beta S}, & t \in [t_i, t_i + f\tau) \\ 0, & t \in [t_i + f\tau, t_i + \tau] \end{cases} \tag{9}$$

For some value of $t_i \in \mathbb{R}$ and $f \in (0, 1]$. We divide the proof into a series of lemmas.

Lemma 3: More fire, less fuel. *Let $b_x(t)$ and $b_y(t)$ be two interventions beginning at the same t_i and lasting until $t_f = t_i + \tau$. Let $I_x(t)$ and $S_x(t)$ be $I(t)$ and $S(t)$, respectively, for the SIR model under intervention $b_x(t)$. Let $I_y(t)$ and $S_y(t)$ likewise by $I(t)$ and $S(t)$ for the SIR under intervention $b_y(t)$. If $I_x(t) \geq I_y(t)$ for all $t \in [t_i, t_f]$, and $I_x(t) > I_y(t)$ for some $t \in (t_i, t_f)$, then $S_x(t_f) < S_y(t_f)$.*

Proof. The difference $\Delta R_{xy}(t) = R_x(t) - R_y(t)$ obeys $\frac{d\Delta R_{xy}}{dt} = \gamma(I_x - I_y)$, which is non-negative for all $t \in [t_i, t_f]$. Therefore $\Delta R_{xy}(t)$ is non-decreasing during any time interval contained in $[t_i, t_f]$. Take $\bar{t} \in (t_i, t_f)$ such that $I_x(\bar{t}) - I_y(\bar{t}) > 0$. Because the intervention $b(t)$ allows only a finite number of discontinuities, the resulting $I(t)$ is continuous. And so there is an ϵ such that $I_x(t) - I_y(t) > 0$ for all $t \in [\bar{t} - \epsilon, \bar{t} + \epsilon]$. Therefore $\Delta R_{xy}(t)$ must be strictly increasing during $[\bar{t} - \epsilon, \bar{t} + \epsilon]$.

Because $\Delta R_{xy}(t)$ never decreases and sometimes increases during $[t_i, t_f]$, it follows that $\Delta R_{xy}(t_i) < \Delta R_{xy}(t_f)$, but $\Delta R_{xy}(t_i) = 0$ so $\Delta R_{xy}(t_f) > 0$ and $R_x(t_f) > R_y(t_f)$. But $I_x(t_f) \geq I_y(t_f)$ and $S = 1 - (I + R)$, so $S_x(t_f) < S_y(t_f)$. □

Lemma 4: Do nothing for a bigger blaze. *Let $b_1(t) \equiv 1$ be the null intervention and take t_{max} such that $I_1(t_{max}) = I_1^{max}$. Let $b_x(t)$ be an intervention that satisfies $b_x(t) < 1$ for almost all $t \in [\bar{t} - \epsilon, \bar{t} + \epsilon]$ for some ϵ and some $\bar{t} < t_{max}$. Then $I_x(t) < I_1(t)$ for all $t \in [\bar{t}, t_{max}]$.*

Proof. Let the time of divergence of the interventions be given by $t_d = \inf\{t \in (t_i, t_{max}) \mid b_x(t) < 1\}$. Note that this infimum must exist because it is taken over a bounded and non-empty set. It is easy to see that $b_x(t) = 1$ and that $I_x(t) = I_1(t)$ for all $t < t_d$. Also, by the right continuity of b_x, there exists an interval $(t_d, t_d + \epsilon)$ such that $I_1 > I_x$ and $b_1 \beta S_1 I_1 > b_x \beta S_x I_x$. Suppose there exists a minimal $t_e \in (t_d, t_{max})$ such that $I_x(t_e) = I_1(t_e)$. We will find a contradiction. Note that $I_x < I_1$ in (t_d, t_e). Because I_1 is monotonic and continuous in $(-\infty, t_{\max})$, we can invert I_1 and define $\hat{t} = I_1^{-1}(I_x(t))$ for all $t \in (-\infty, t_{\max})$.

We can see that $\frac{dI_1}{dt}(\hat{t})$ is the growth rate of the infectious class under the null intervention when $I_1 = I_x(t)$. It follows that $\frac{dI_1}{dt}(\hat{t}) < \frac{dI_x}{dt}(t)$ for some value of t, otherwise I_1 would always dominate I_x, and t_e would not exist. Let $t_{\inf} = \inf\{t \in (t_d, t_e) \mid \frac{dI_1}{dt}(\hat{t}) < \frac{dI_x}{dt}(t)\}$. This implies that $S_1(\hat{t}) \leq S_x(t)$ for some interval including $t_{\inf}$ that can be maximally extended to the left as $(t_{\min}, t_{\inf}]$. By continuity of S, $S_1(\hat{t}_{\min}) = S_x(t_{\min})$.

If $t_{\min} \neq t_d$, then $t_{\min} > \hat{t}_{\min}$, and $R_1(\hat{t}_{\min}) = R_x(t_{\min})$, but this is impossible because by construction $\frac{dI_1}{dt}(\hat{t}) > \frac{dI_x}{dt}(t)$ for all $t \in (t_d, t_{\min})$, and therefore $R_1(\hat{t}_{\min}) < R_x(t_{\min})$.

If $t_{\min} = t_d$, then $t_{\min} = \hat{t}_{\min}$. Also, $R_1(\hat{t}) > R_x(t)$ for all $t \in (t_d, t_{inf})$. But that is impossible because $\frac{dR_1(\hat{t}_d)}{dt} = \frac{b(t_d)\beta SI - \gamma I}{\beta SI - \gamma I}\gamma I < \gamma I = \frac{dR_x(t_d)}{dt}$, and trivially $R_1(\hat{t}_d) = R_x(t_d)$. □

Lemma 5: Wait, maintain, suppress. *Let b_x be any intervention, and $I_x^{max} = \max\{I_x(t)\}$. There exists an intervention b_y with same beginning and duration as b_x of the form*

$$b_y(t) = \begin{cases} 1 & t \in [t_i, t_i + g\tau) \\ \frac{\gamma}{\beta S}, & t \in [t_i + g\tau, t_i + f\tau) \\ 0, & t \in [t_i + f\tau, t_i + \tau] \end{cases} \tag{10}$$

For some $g < f$ and $g, f \in (0, 1]$, such that $I_y^{max} < I_x^{max}$.

Proof. Define $I_x^\tau = \max\{I_x(t) \mid t \in [t_i, t_f]\}$. Following lemma 4, take g such that $t_i + g\tau = I_1^{-1}(I_x^\tau)$. Now take f such that $(1 - f)\tau = \frac{1}{\gamma} \log \frac{I_x^\tau}{I_x(t_f)}$. From lemma 4 it is clear that $I_y(t) \geq I_x(t)$ for $t \in [t_i\tau, t_i + g\tau)$. By construction $I_y(t) = I_x^\tau$ for $t \in [t_i + g\tau, t_i + f\tau)$, and therefore the inequality remains true. Finally, for $t \in [t_i + f\tau, t_i + \tau]$, because $I_y(t)$ decays faster than $I_x(t)$, then if $I_y(t) < I_x(t)$ at any time, then $I_y(t_f) < I_x(t_f)$, but by construction $I_y(t_f) = I_x(t_f)$. By lemma 3, $S_y(t_f) < S_x(t_f)$, and by lemma 2, $I_y^{max} < I_x^{max}$.

It is possible however that this proposed b_y is not a viable intervention if $S_y(t_y^{crit}) = S_{crit}$ for some $t_y^{crit} \in [t_i + g\tau, t_i + f\tau)$, as this would require b_y to assume values larger than 1. If that is the case we can define $b_{\bar{y}}$, by taking f such that $t_i + f\tau = t_y^{crit}$. Note that because by the end of the intervention $b_{\bar{y}}$ the infectious class will monotonically decrease because $S_{\bar{y}} \leq S_{crit}$. This implies that $I_{\bar{y}}^{max} = I_{\bar{y}}^\tau = I_x^\tau < I_x^{max}$. □

Theorem 1: Maintain, then suppress. *An intervention $b_{opt}(t)$ such that $I_{opt}^{max} \leq I_x^{max}$ for any intervention $b_x(t)$ must take the form*

$$b_{opt}(t) = \begin{cases} \frac{\gamma}{\beta S}, & t \in [t_i, t_i + f\tau) \\ 0, & t \in [t_i + f\tau, t_i + \tau] \end{cases} \tag{11}$$

For some value of $t_i \in \mathbb{R}$ and $f \in (0, 1]$.

Proof. If $b_x(t)$ is an optimal intervention, then lemma 5 insures that $b_x(t)$ is of the form given by equation 10. Now consider the strategy $b_{opt}(t)$ of the form given by equation 3 with $t_i^{opt} = t_i^x + g\tau$ and $f^{opt} = f^x - g$. This new intervention functions exactly like b_x for the entire duration of b_x, but then b_{opt} is held at 0 for a little longer further reducing I_{opt}. This means that $S_{opt}(t_f^{opt}) = S_x(t_f^x)$ and $I_{opt}(t_f^{opt}) \leq I_x(t_f^x)$ and therefore, by lemma 1, $I_{opt}^{max} \leq I_x^{max}$. □

Remark 2: . *Because during an intervention b_{opt} the susceptible class is depleted at a constant rate for all $t \in [t_i, t_i + f\tau)$, b_{opt} can be written as*

$$b_{opt}(t) = \begin{cases} \frac{\gamma}{\beta\left(S_{t_i} - I_{t_i}\gamma(t - t_i)\right)}, & t \in [t_i, t_i + f\tau) \\ 0, & t \in [t_i + f\tau, t_i + \tau] \end{cases} \tag{12}$$

For $S_{t_i} = S(t_i)$ and $I_{t_i} = I(t_i)$.

B.4 Results of an optimal intervention

It follows that, given an optimal intervention b of duration τ and depletion fraction f begun at time t_i and ending at $t_f = t_i + \tau$:

$$S(t_f) = S(t_i) - \gamma\tau f I(t_i) \tag{13}$$

$$I(t_f) = I(t_i) \exp\left[-\gamma\tau(1 - f)\right] \tag{14}$$

In an SIR system without intervention that begins with a wholly susceptible population, we have the relation:

$$I(S) = 1 - S + \frac{1}{\mathcal{R}_0} \log(S) \tag{15}$$

And so:

$$I(t_i) = 1 - S(t_i) + \frac{1}{\mathcal{R}_0} \log(S(t_i)) \tag{16}$$

And we can likewise write $S(t_f)$ and $I(t_f)$ in terms of $S(t_i)$:

$$S(t_f) = S(t_i) - \gamma\tau f\left(1 - S(t_i) + \frac{1}{\mathcal{R}_0} \log(S(t_i))\right) \tag{17}$$

$$I(t_f) = \left(1 - S(t_i) + \frac{1}{\mathcal{R}_0}\log(S(t_i))\right)\exp\left[-\gamma\tau(1-f)\right] \tag{18}$$

This in turn allows us to express $I^{\max}(t_f)$ purely in terms of $S(t_i)$ and the parameters, though the expression is long:

$$\begin{aligned} I^{\max}(t_f) = &\left(1 - S(t_i) + \frac{1}{\mathcal{R}_0}\log(S(t_i))\right)\exp\left[-\gamma\tau(1-f)\right] \\ &+ S(t_i) - \gamma\tau f\left(1 - S(t_i) + \frac{1}{\mathcal{R}_0}\log(S(t_i))\right) \\ &- \frac{1}{\mathcal{R}_0}\log\left(S(t_i) - \gamma\tau f\left(1 - S(t_i) + \frac{1}{\mathcal{R}_0}\log(S(t_i))\right)\right) \\ &- \frac{1}{\mathcal{R}_0} + \frac{1}{\mathcal{R}_0}\log\left(\frac{1}{\mathcal{R}_0}\right) \end{aligned} \tag{19}$$

Remark 3: Continuity of $I(t_i), I^{\max}(t_f)$. *Notice that both $I(t_i)$ and $I^{max}(t_f)$ are continuous functions of $S(t_i)$, since they are both linear combinations of continuous functions of $S(t_i)$ and since $S(t)$ is continuous in t_i, they are continuous functions of t_i.*

Lemma 6: Don't be late. *Let t_p be the infimum of times such that $I(t) = I^{max}$, and let t_i be the start of an optimal intervention. Then $t_p \geq t_i$. That is, I^{max} cannot occur before the intervention begins for an optimal intervention.*

Proof. Since $b(t) = 1$ for $t < t_i$ $t_p < t_i$ necessarily implies that $S(t_i) < S_{\text{crit}}$, that is, the epidemic is already declining when the intervention begins and will never grow again, regardless of the intervention approach (since $b(t) \leq 1$, we cannot force a declining epidemic to grow). That in turn implies $I^{\max} = I^{\text{peak}}$, which we can with certainty improve upon. So $I^{\max}$ must occur during or after the intervention. □

Corollary 1: Start with fuel. *It is immediate that $S(t_i) > S_{crit}$.*

Corollary 2: Peak early. *Since $I(t) \leq I(t_i)$ for $t \in [t_i, t_i + \tau]$ during an optimal intervention, if I^{max} occurs during the intervention, $I^{max} = I(t_i)$*

Theorem 2: Twin Peaks. *Let $b_x(t)$ be an optimal intervention, then $I_x(t) \leq I_x^{max}$ for all $t \in (-\infty, t_f]$, with equality for $t \in [t_i, t_i + \tau f]$, and furthermore $I_x(t_p) = I_x^{max}$ for some $t_p \in [t_f, \infty)$ with $t_p = t_f$ only if $f = 1$.*

That is, if $0 < f < 1$, there will be a plateau during the intervention followed by a peak of equal height that occurs strictly after the intervention finishes. If $f = 0$, there will be two peaks of equal height, one at the start of the intervention and one strictly after it finishes, and if $f = 1$ there will be a plateau during the intervention with no subsequent peak.

Proof. Let b_x be an optimal intervention of the form given by equation 3. By Lemma 6, I_x^{max} must occur during or after the intervention. First let us assume that the I_x^{max} occurs during the intervention and is never again attained after the intervention. If $f = 1$, this implies that the whole intervention is a plateau and no further peaks occur. For $f < 1$ we can build a new intervention b_y of the same form as b_x but that starts at $t_i - \epsilon$ rather than at t_i. From the continuity of I^{max} in t_i for any intervention, it follows that for some ϵ small enough, I_y^{max} post intervention must still be smaller than I_y^{max} during the intervention, but because $I(t_i - \epsilon) < I(t_i)$, corollary 2 implies that our new b_y outperforms b_x, which contradicts the optimality of b_x.

Now let us assume that $I_x^{\max}$ occurs after the intervention and is larger than any value of I_x during the intervention. If $f > 0$, we can once again build a new intervention function b_y

$$b_y(t) = \begin{cases} 1 & t \in [t_i, t_i + \epsilon) \\ \frac{\gamma}{\beta S}, & t \in [t_i + \epsilon, t_i + f_y\tau) \\ 0, & t \in [t_i + f_y\tau, t_i + \tau] \end{cases} \tag{20}$$

With f_y chosen such that $(1-f_y)\tau = \frac{1}{\gamma}\log\frac{I_x(t_i+\epsilon)}{I_x(t_f)}$. For sufficiently small ϵ, $I_y(t_i+\epsilon) < I_x^{\max}$. Also, following lemma 5, $I_y^{\max} < I_x^{\max}$ and therefore b_y outperforms b_x, which once again contradicts the optimality of b_x.

Finally, if $f = 0$, we build yet another b_y of the same form as b_x but that starts at $t_i + \epsilon$ rather than at t_i. Because intervention b_y starts out with a smaller susceptible fraction than intervention b_x, and an increase in the infected fraction that is smaller than the susceptible fraction decrease, it follows from equation 6 that $I_y^{\max} < I_x^{\max}$ and therefore b_y outperforms b_x, which once again contradicts the optimality of b_x. □

Corollary 3: If you don't have much gunpowder, don't shoot until you see the whites of their eyes. *Given the optimal intervention of duration τ referred to as b_{opt}^{τ}, let the initial time of such an intervention be referred to as t_i^{τ}. As $\tau \to 0$, $t_i^{\tau} \to t_{crit}$.*

Proof. From equation 19 we can see that I_{opt}^{max} is a continuous function of τ, and that as $\tau \to 0$, the effect of intervention defined as $I_1^{max} - I_{opt}^{max} \to 0$. We know from theorem 2 that $I_{opt}(t_i) = I_{opt}^{max}$, and therefore, as $\tau \to 0$, $I_{opt}(t_i) \to I_1^{max}$, which implies $t_i^{\tau} \to t_{crit}$. □

Corollary 4: No need to burn all the fuel. *After an optimal intervention, $S(t_f) \geq S_{crit}$, with equality if and only if $f = 1$.*

Proof. By Theorem 2, $I(t_i) = I^{\max}(t_f) = I^{\max}$. If $f < 1$, we must have $I(t_f) < I(t_i) = I^{\max}$, so in order for the epidemic to reach $I^{\max}(t_f) = I^{\max}$, we must have $S(t_f) > S_{\text{crit}}$. If $f = 1$, then $I(t_f) = I(t_i) = I^{\max}$, so then must have $S(t_f) = S_{\text{crit}}$, otherwise the epidemic would grow to a peak above $I(t_i) = I(t_f)$, which would be a contradiction of Theorem 2. □

Corollary 5: Putting out existing fire can only do so much. *Consider a full suppression intervention of duration τ defined by $b_0(t) = 0$ for all $t \in [t_i, t_i + \tau]$. For every τ there is a t_i that minimizes I_0^{max}. Consider these optimized full suppression interventions. Then, as $\tau \to \infty$, the maximum infectious prevalence $I_0^{max} \to \frac{1}{2} + \frac{1}{2\mathcal{R}_0}\left(\log\left(\frac{1}{\mathcal{R}_0}\right) - 1\right)$. In other words, full suppression interventions have a limit in how much they can reduce I^{max}.*

Proof. From the proof of theorem 2 for $f = 0$, it follows that full suppression interventions have an optimal start time t_i, and that $I_0(t_i) = I_0^{\max}$. Also, because no new infections occur during a full suppression intervention, $S_0(t_i) = S_0(t_f)$ so substituting into equation 6

$$I_0^{\max}(t_i) = I_0(t_f) + S_0(t_i) - \frac{1}{\mathcal{R}_0}\log\Big(S_0(t_i)\Big) - \frac{1}{\mathcal{R}_0} + \frac{1}{\mathcal{R}_0}\log\left(\frac{1}{\mathcal{R}_0}\right) \tag{21}$$

We the further use equation 15 to substitute $S_0(t_i)$ and take $\tau \to \infty$ such that $I_0(t_f) \to 0$ which finally yields

$$I_0^{\max} = \frac{1}{2} + \frac{1}{2\mathcal{R}_0}\Big(\log\Big(\frac{1}{\mathcal{R}_0}\Big) - 1\Big) \tag{22}$$

□

Corollary 6: But when in doubt, put out the fire. *Let b_0 be the optimized full suppression intervention with duration τ and starting time t_i^0. For a full suppression intervention $b_{\hat{0}}$ that also has duration τ but that has a starting time $t_i^{\hat{0}} \in [t_i^0, t_{crit}]$, the infected fraction peak is $I_{\hat{0}}^{max} = I_{\hat{0}}(t_i^{\hat{0}})$, moreover no intervention starting at that same time can attain a lower peak.*

Proof. It suffices to show that delaying a full suppression intervention by an infinitesimal ϵ diminishes the secondary peak. From equation 24 with $f = 0$, it is clear that

$$\begin{aligned}
\frac{\partial}{\partial t_i} I^{\max}(t_f) &= -\frac{S'(t_i)}{\mathcal{R}_0 S_i} + e^{-\gamma\tau} I'(t_i) + S'(t_i) \\
\frac{\partial}{\partial t_i} I^{\max}(t_f) &= \gamma I_i + e^{-\gamma\tau}(\beta I_i S_i - \gamma I_i) - \beta I_i S_i \\
\frac{\partial}{\partial t_i} I^{\max}(t_f) &= (e^{-\gamma\tau} - 1)(I'(t_i))
\end{aligned} \tag{23}$$

But $I'(t_i) > 0$ for $t_i < t_{crit}$ and $(e^{-\gamma\tau} - 1) < 0$ which means that delaying the full suppression decreases the post intervention peak. Trivially no intervention can attain a peak lower than its initial condition, which concludes the proof. □

B.5 Optimization of t_i, f

Applying equations 13, 14, the partial derivatives of $I^{\max}(t_f)$ with respect to f and t_i are given by:

$$\begin{aligned}
\frac{\partial}{\partial f} I^{\max}(t_f) &= \frac{(\gamma\tau I_i)}{\mathcal{R}_0\Big(S_i - \gamma\tau f\Big) I_i} + \gamma\tau I_i e^{-(1-f)\gamma\tau} - \gamma\tau I_i \\
\frac{\partial}{\partial t_i} I^{\max}(t_f) &= -\frac{S'(t_i) - f\gamma\tau I'(t_i)}{\mathcal{R}_0\Big(S_i - f\gamma\tau I_i\Big)} + e^{-(1-f)\gamma\tau} I'(t_i) - f\gamma\tau I'(t_i) + S'(t_i)
\end{aligned} \tag{24}$$

Substituting the values of $\frac{\mathrm{d}S}{\mathrm{d}t}$, $\frac{\mathrm{d}I}{\mathrm{d}t}$:

$$\begin{aligned}
\frac{\partial}{\partial t_i} I^{\max}(t_f) = &-\frac{(-\beta S_i I_i) - f\gamma\tau\,(\beta S_i I_i - \gamma I_i)}{\mathcal{R}_0\,(S_i - f\gamma\tau I_i)} \\
&+ e^{-(1-f)\gamma\tau}\,(\beta S_i I_i - \gamma I_i) \\
&- f\gamma\tau\,(\beta S_i I_i - \gamma I_i) - \beta S_i I_i
\end{aligned} \tag{25}$$

Setting equal to zero and simplifying:

$$\frac{\partial}{\partial f} I^{\max}(t_f) = \frac{1}{\mathcal{R}_0\Big(S_i - f\gamma\tau I_i\Big)} + \gamma\tau e^{-(1-f)\gamma\tau} - 1 = 0 \tag{26}$$

$$\begin{aligned}\frac{\partial}{\partial t_i} I^{\max}(t_f) = &-\frac{(-\beta S_i) - f\gamma\tau\left(\beta S_i - \gamma\right)}{\mathcal{R}_0\left(S_i - f\gamma\tau I_i\right)} + e^{-(1-f)\gamma\tau}\left(\beta S_i - \gamma\right) \\ &- f\gamma\tau\left(\beta S_i - \gamma\right) - \beta S_i = 0\end{aligned} \tag{27}$$

$$\begin{aligned}\frac{\partial}{\partial S} I^{\max}(S(t_i), f) = &\left(\frac{1}{\mathcal{R}_0 S} - 1\right)\left(\exp\left[-\gamma\tau(1-f)\right] - \gamma\tau f\right) \\ &- \left(\frac{1}{\mathcal{R}_0}\right)\frac{1 - f\gamma\tau(\frac{1}{\mathcal{R}_0 S} - 1)}{S - \gamma\tau f(1 - S + \frac{1}{\mathcal{R}_0}\log S)} + 1\end{aligned} \tag{28}$$

$$\begin{aligned}\frac{\partial}{\partial f} I^{\max}(S(t_i, f)) = &\left(\frac{1}{\mathcal{R}_0 S} - 1\right)\exp\left[-\gamma\tau(1-f)\right] \\ &+ 1 - \gamma\tau f\left(\frac{1}{\mathcal{R}_0 S} - 1\right) \\ &- \left(\frac{1}{\mathcal{R}_0}\right)\frac{\frac{1}{\mathcal{R}_0 S} - 1}{S - \gamma\tau f(1 - S + \frac{1}{\mathcal{R}_0}\log S)}\end{aligned} \tag{29}$$

Lemma 7: If you have lots of gunpowder, shoot early. *For an optimal intervention b^{τ}_{opt} acting on a SIR with full susceptibility as initial condition, as $\tau \to \infty$, $S^{\tau}_{opt}(t_i) \to 1$; that is, we intervene almost immediately, when almost all the population is fully susceptible.*

Proof. Define an auxiliary intervention b^{τ}_x with $\tau > \frac{1-S_{crit}}{\gamma I^{max}}$ such that

$$b^{\tau}_x(t) = \begin{cases} \frac{\gamma}{\beta S}, & \text{if } t > t_x \text{ and } S > S_{crit} \\ 1, & \text{elsewhere} \end{cases} \tag{30}$$

with $t_x = I_1^{-1}(\frac{1-S_{crit}}{\gamma\tau})$. It is clear that such an intervention has a duration of τ or less, and that by the end of such an intervention $S_x \leq S_{crit}$. Therefore $I^{max}_x = I_x(t_x) = \frac{1-S_{crit}}{\gamma\tau}$, but by definition $I^{max}_{opt} \leq I^{max}_x$. Combining with corollary 2, $I^{\tau}_{opt}(t_i) \leq \frac{1-S_{crit}}{\gamma\tau}$ so as $\tau \to \infty$ both $I^{\tau}_{opt}(t_i) \to 0$ and $S^{\tau}_{opt}(t_i) \to 1$. □

Theorem 3: Always maintain, always suppress. *For an SIR with full susceptibility as initial condition, it is the case that any optimal intervention with positive duration as defined by equation 3 has $0 < f < 1$. In other words, the optimal intervention for an emerging pathogen ($S(0) \to 1$, $I(0) \to 0$) always has a maintenance phase and always has a total suppression phase.*

Proof. Suppose $f = 1$. From equation 26,

$$\frac{1}{\mathcal{R}_0\Big(S_i - \gamma\tau I_i\Big)} + \gamma\tau - 1 = 0 \tag{31}$$

But following corollary 4, $S_i - \gamma\tau I_i = S_{crit}$, which substituting in the previous equation implies that $\gamma\tau = 0$, which is impossible. Therefore f cannot be 1.

Now let us assume $f = 0$. From equation 27 we have that

$$\begin{aligned} \gamma + e^{-\gamma\tau}(\beta S_i - \gamma) - \beta S_i = 0 \\ (e^{-\gamma\tau} - 1)(\beta S_i - \gamma) = 0 \end{aligned} \tag{32}$$

Which implies that either $(e^{-\gamma\tau} - 1) = 0$, and therefore $\tau = 0$, or $(\beta S_i - \gamma) = 0$ and $S_i = S_{crit}$ which also implies $\tau = 0$, which is impossible. □

B.6 A general classification of interventions

From equation 6 it is clear that there are two fundamental methods of reducing the peak of an epidemic: depleting the infected fraction and depleting the susceptible fraction. We have shown that different interventions achieve peak reduction with different combinations of those methods. Our optimal intervention, for example, is characterized by a pure susceptible depletion phase followed by a pure infected depletion phase. Full suppression interventions, in contrast, operate solely by depleting the infected fraction. We observe that interventions can be classified in terms of how much they rely on depleting the susceptible fraction versus depleting the infected fraction.

The effect of an intervention can be understood as the infectious peak if no intervention were to take place minus the infectious peak given the intervention. In a more formal notation, an intervention b_x has an effect $I^{\max} - I_x^{\max} = I^{\max}(I_x(t_i), S_x(t_i)) - I^{\max}(I_x(t_f), S_x(t_f)) = \Delta_x(t_f)$. By applying equation 6 we obtain

$$\begin{aligned} \Delta_x(t_f) &= -\Big[I_x(t_f) - I_x(t_i)\Big] - \Big[G\Big(S_x(t_f)\Big) - G\Big(S_x(t_i)\Big)\Big] \\ G(S) &= S - \frac{1}{\mathcal{R}_0}\log(S) \end{aligned} \tag{33}$$

Then if $-\Big[I_x(t_f) - I_x(t_i)\Big] > -\Big[G\Big(S_x(t_f)\Big) - G\Big(S_x(t_i)\Big)\Big]$ we can say that the intervention b_x overall relies more on infected depletion, whereas if the opposite is true we can say that it relies more on susceptible depletion. Moreover, by applying the fundamental theorem of calculus, we obtain

$$\int_{t_i}^{t_f} \Delta_x'(t)dt = -\int_{t_i}^{t_f} I_x'(t) + S_x'(t)\Big(1 - \frac{1}{\mathcal{R}_0 S_x(t)}\Big)dt \tag{34}$$

Which allows us to look at a certain time $t \in [t_i, t_f]$ and say that if

$$I_x'(t) < S_x'(t)\Big(1 - \frac{1}{\mathcal{R}_0 S_x(t)}\Big) \tag{35}$$

Then at that moment t, the intervention b_x acts more by depleting the infected fraction than by depleting the susceptible fraction. The condition can be simplified to

$$b_x(t)\Big(2\mathcal{R}_0 S_x(t) - 1\Big) < 1 \tag{36}$$

References

[1] Helen Branswell. "Why 'flattening the curve' may be the world's best bet to slow the coronavirus". In: *STAT News* (Mar. 2020). URL: https://www.statnews.com/2020/03/11/flattening-curve-coronavirus/.

[2] Francesco Di Lauro, István Z Kiss, and Joel Miller. "The timing of one-shot interventions for epidemic control". In: *medRxiv* (Mar. 2020). DOI: 10.1101/2020.03.02.20030007.

[3] Neil M. Ferguson et al. *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*. Imperial College, London, Mar. 2020. URL: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/.

[4] Barney S. Graham, John R. Mascola, and Anthony S. Fauci. "Novel vaccine technologies: essential components of an adequate response to emerging viral diseases". In: *Jama* 319.14 (2018), pp. 1431–1432.

[5] World Health Organization Writing Group. "Nonpharmaceutical interventions for pandemic influenza, national and community measures". In: *Emerging infectious diseases* 12.1 (2006), p. 88.

[6] William Ogilvy Kermack and Anderson G. McKendrick. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pp. 700–721.

[7] Stephen M. Kissler et al. "Social distancing strategies for curbing the COVID-19 epidemic". In: *medRxiv* (Mar. 2020). DOI: 10.1101/2020.03.22.20041079.

[8] Ruiyun Li et al. "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)". In: *Science* (2020).

[9] Sang Woo Park et al. "Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (SARS-CoV-2) outbreak". In: *medRxiv* (2020).

[10] Michael D. Shear et al. "The Lost Month: How a Failure to Test Blinded the U.S. to Covid-19". In: *The New York Times* (Mar. 2020), p. 1. URL: https://www.nytimes.com/2020/03/28/us/testing-coronavirus-pandemic.html.

[11] Howard (Howie) Weiss. "The SIR model and the foundations of public health". In: *MATerials MATemàtics* (3 2013), pp. 0001–17.

[12] Lirong Zou et al. "SARS-CoV-2 viral load in upper respiratory specimens of infected patients". In: *New England Journal of Medicine* 382.12 (2020), pp. 1177–1179.