Road to recovery: Managing an epidemic*

Simon Loertscher† Ellen V. Muir‡

This version: January 3, 2021 First version: April 23, 2020

Abstract

Without widespread immunization, the road to recovery from the current COVID-19 lockdowns will optimally follow a path that finds the difficult balance between the social and economic benefits of liberty and the toll from the disease. We provide an approach that combines epidemiology and economic models, taking as given that the maximum capacity of the healthcare system imposes a constraint that must not be exceeded. Treating the transmission rate as a decreasing function of the severity of the lockdown, we first determine the minimal lockdown that satisfies this constraint using an epidemiology model with a homogeneous population to predict future demand for healthcare. Allowing for a heterogeneous population, we then derive the optimal lockdown policy under the assumption of homogeneous mixing and show that it is characterized by a bang-bang solution. Possibilities such as the capacity of the healthcare system increasing or a vaccine arriving at some point in the future do not substantively impact the dynamically optimal policy until such an event actually occurs.

Keywords: COVID-19, SIR models, capacity constraints, managing an epidemic
JEL-Classification: H51, I18

*We thank an anonymous referee and the co-editors of this journal for comments and suggestions that have helped us improve the paper. We are also grateful for feedback from and discussion with Eric Budish, Gabriel Carroll, Chris Edmond, Ian Harper, Zi Yang Kang, Paul Milgrom, Martin Souchier, Gary Stoneham and seminar audiences at the COVID-19 Policy Hackathon organized by the Stanford Economic Association and the MIT Undergraduate Economic Association and at the University of Melbourne. Anand Bharadwaj provided excellent research assistance. Financial support by the June and Samuel Hordern Endowment is also gratefully acknowledged.

†Department of Economics & Centre for Market Design, Level 4, FBE Building, 111 Barry Street, University of Melbourne, Victoria 3010, Australia. Email: simonl@unimelb.edu.au.

‡Department of Economics, Stanford University. Email: evmuir@stanford.edu. Corresponding author.

1 Introduction

Without widespread immunization of the population, the road to recovery from pandemic-induced lockdowns requires sustained vigilance to ensure that the spread of the disease remains at a level that is manageable for a country's or region's healthcare system. At the same time, recovery ought to start as soon as possible to limit the reduction in liberty that such lockdowns impose, the mental and other health issues associated with social distancing and isolation, and to minimize the economic cost. If eradication is impossible or possible only at tremendous costs, keeping the pandemic under control without inducing economic and social hardship at a catastrophic scale requires finding a path through territory that is uncharted for both epidemiologists and economists. From a public health perspective, recovery requires the transition from a paradigm in which eradication of an epidemic is the goal to one in which the epidemic is *managed*. For economists, recovery requires ploughing a path through a system whose dynamics are non-linear.

In this paper, we show how this can be done by providing a methodology that permits the return to some kind of normalcy, while keeping the spread of the disease at a level that even at the peak of the epidemic does not exceed the capacity constraint of the healthcare system. Specifically, we use a standard epidemiology model—a simple *SIR model*—to predict the peak of the epidemic and treat the rate of transmission of the disease as the variable that the policymaker can influence by choosing the severity of a lockdown. We treat as a hard constraint the capacity of the healthcare system, that is, the maximum number of patients that it can handle at the peak of the crisis.[1] Of course, this capacity constraint will need to be defined in such a way that patients with other—but no less severe—needs for care are still able to access treatment.[2]

Our framework resonates with recent policy, as the capacity constraints-based approach allows policymakers to avoid explicitly trading off dollars against lives.[3] In the U.S., as

[1]To fix ideas, throughout the paper we will talk about the capacity of the healthcare system as our binding constraint. However, the framework outlined in this paper can accommodate any constraint that can be expressed as a function of the number of cases that occur at the peak of the epidemic. This constraint can be interpreted as a normative criterion that society has to choose. For example, suppose that society viewed the number of deaths that would occur as a result of using the capacity of the healthcare system as a binding constraint as unacceptable. Then one could instead treat a fixed proportion of the healthcare system that is utilized at the height of the crisis as a binding constraint (such as requiring that the healthcare system never exceeds 80% capacity). An alternative, but mathematically equivalent, approach would be to place an upper bound on the number of deaths per day at the peak of the crisis, which may be in line with some of the concurrent debates (see, for example, the New York Times article "The Cold Calculations America's Leaders Will Have to Make Before Reopening" (April 22, 2020)).

[2]In a public health catastrophe, this is not always the case; see, for example, this New York Times article: "The Pandemic's Hidden Victims: Sick or Dying, but Not From the Virus" (April 20, 2020).

[3]As New York Governor Andrew Cuomo put it in May 2020: "To me, I say the cost of a human life is priceless, period. Our reopening plan doesn't have a trade-off." (Source: Buffallonews.com, May 5, 2020)

they approach the winter peak of the pandemic, the states of California and New York have both adopted policies that make lockdowns—and more generally restrictions on public life—contingent on hospital capacity utilization. For example, starting in December 2020, any region in California goes into lockdown as soon as the available ICU hospital capacity dips below 15%.[4] In New York, the "state's new approach focuses on maintaining sufficient hospital capacity instead of shutting down economic activity," according to the New York Times article "Cuomo Tries to Jolt Public by Warning of Overwhelmed Hospitals" (December 12, 2020). Moreover, this new policy involves an element of prediction-based restrictions, precisely as implied by our approach: According to the same article, the "most complex element, which could prompt regionwide shutdowns, involves taking the rate of increase in an area's hospitalizations and projecting forward to determine whether it would top 90% of capacity in three weeks. If so, restrictions will be introduced that include the closing of nonessential businesses, the limiting restaurants to takeout and delivery [sic] and a prohibition on nearly all gatherings."

The main contribution of this paper is to formulate an operational constraint that provides policymakers with guidance for how to manage an epidemic which is too costly to eradicate, to incorporate this constraint into a standard epidemiology model, and to determine the severity of the lockdown that is necessary to respect the constraint. Allowing for heterogeneity in the population, we show that managing an epidemic subject to a capacity constraint involves a non-trivial economic optimization problem without requiring the policymaker to take a stance on the value of life because optimality requires satisfying the constraint at minimum economic cost. Extending our model to consider dynamically optimal policies, we show that possibilities such as the capacity of the healthcare system increasing or a vaccine arriving at some point in the future do not substantively impact the optimal policy until such an event actually occurs. While the purpose of our model is to serve as a proof of concept that would need to be refined if applied, many of the key insights—such as the need to use epidemiology models to predict future healthcare demand and the non-trivial economic optimization problem when faced with a capacity constraint—will extend well beyond the confines of the specific setups we study.

The economic toll from the lockdowns implemented due to the COVID-19 pandemic has little parallel in living memory. To convey a sense of the magnitude of the potential economic and social costs, consider the unemployment rate during the Great Depression in the U.S.,

Of course, we are not to suggesting that there are no tradeoffs, but rather that in the environment in which public policy and political debates take place it is useful if these tradeoffs are derived from capacity constraints that, at any point in time, are largely a given for the decision makers.

[4]See the San Francisco Chronicle artile "How Bay Area ICU capacity compares to the most impacted areas in California, nation" (December 11, 2020).

then and now the world's largest economy, and the unemployment rates before and in the wake of the ongoing lockdowns in the U.S. in 2020. The immediate consequences of the Great Depression were mass poverty and economic devastation, and at least indirectly, the rise of fascism in Europe. As Table 1 shows, the unemployment rate in the U.S. rose sharply from 5.2% in March 2020 to 19.5% in April 2020 as the nationwide lockdown hit the country and much of the rest of the world economy, and then steadily declined as the economy began to reopen. This steep incline and swift partial recovery reflects the peculiarity of the present economic downturn, which was *not* caused by a bad state of the economy. This is at the same time a source of hope and of concern: while the healthy underlying state of the economy at the onset may make for a relatively fast recovery, extended or repeated complete lockdowns can turn a public health shock into a deep and prolonged economic crisis. The firms that workers could return to in May and June may simply go out of business after further or extended lockdowns. Thus, the problem of finding a smooth path to recovery is particularly salient.

Great D.	1929	1930	1931	1932	1933
Unemployment rate[5]	3.2	8.7	15.9	23.6	24.9

2020	Mar 20	Apr 20	May 20	Jun 20
Unemployment rate[6]	5.2	19.5	16.4	11.1

Table 1: Upper table: unemployment rates during the Great Depression in the U.S. Lower table: weekly unemployment filings (in thousands) in the U.S. in 2020.

As documented by Brodeur et al. (2020), who counted 106 NBER publications over a ten week period in March and April of 2020, there has been an upsurge of interest in the economics of COVID-19 that coincided with the first wave of the pandemic hitting Europe and North-America.[7] Atkeson (2020) provides an introduction for economists to the SIR modeling approach, which is standard in mathematical biology (see, for example, Murray, 2002). As our paper utilizes the SIR framework, it is most closely related to other SIR-based economics papers, including Alvarez et al. (2020), Acemoglu et al. (2020), and Farboodi et al. (2020), who derive optimal lockdown policies for a planner that assigns some weight to economic output and some weight to human life.[8] Alvarez et al. (2020) apply an optimal

[5]Sources: thebalance.com and Reinhart and Rogoff (2009).

[6]Source: US Bureau of Labor. For March, April and May, the table displays the rates that are adjusted for a counting error. With the adjustments, the respective rates would be 4.4, 14.7 and 13.3. For June, the official rate is displayed as an adjusted rate was not available.

[7]There has also been a large volume of more informal commentary on and discussions of the problems at hand (see, for example, Gilbert et al., 2020) and analyses of tradeoffs involving economics without explicitly embedding epidemiology models such as Budish (2020). Stock (2020) provides a short survey.

[8]This objective is similar to that considered in Hall et al. (2020), who study the tradeoff between con-

control approach to an SIR model to derive the optimal lockdown policy that trades off the cost of death against economic output. In this framework, the intensity of the optimal lockdown naturally depends on how the fatality rate varies with the number of infected individuals and the value of a statistical life. These authors also find that the possibility that the pandemic will become easier to manage in the future creates a dynamic complementarity, where the planner has an incentive to induce a stronger lockdown, delaying the spread of the disease until the planner is better equipped to handle it. Acemoglu et al. (2020) provide a multi-group SIR model in which infection, hospitalization and fatality rates vary between groups and find that optimal policies that differentially target these groups outperform non-discriminatory policies. Farboodi et al. (2020) develop a quantitative framework for exploring how individuals trade off the benefits of social activity against the health costs of social activity. They find that the expected cost of COVID-19 in the US is $12,700 per person in a laissez-faire equilibrium and $8,100 per person under an optimal policy. Akbarpour et al. (2020) depart from a classic SIR modeling approach by simulating in an agent-based model calibrated to a rich set of micro-level data to analyze the social, economic and health impacts of alternative policies.

Our paper differs from the aforementioned SIR-based models in that the starting point of our analysis is not that the planner puts one weight on economic output and another on human life. Rather, the optimal policy is derived subject to a capacity constraint, so that the policymaker does not have to take an explicit stance on the value of human life at the outset of the analysis. As mentioned, ensuring that the capacity of the healthcare system is not exceeded is a key factor in preventing catastrophic health outcomes such as those that occurred in the first half of 2020 in Italy's Lombardy region and in New York City. We also find that, under a capacity constraint, when the planner faces the possibilities that the capacity of the healthcare system will increase or a vaccine will arrive at some point in the future there are no dynamic complementarities of the nature explored in Alvarez et al. (2020), as well as many other papers that utilize the same objective function involving lost output and deaths. While we make use of numerical methods to derive optimal dynamic policies, our paper differs methodologically from the previously discussed papers by also deriving analytical results.

The remainder of this paper is organized as follows. Section 2 describes our setup. Section 3 derives the dynamics of an epidemic and the optimal lockdown necessary to keep it at a level that respects the capacity constraint in a homogeneous population model. In Section 4 we derive the optimal lockdown policy for a model with a heterogeneous population and

sumption and lives saved from a utilitarian perspective, assuming that the disease reduces the survival rate of older individuals by a constant for one period.

show that the optimal policy takes a bang-bang form under homogeneous mixing. Section 5 extends the analysis by deriving the dynamically optimal policy in a discrete-time version of the model. It also augments the model by allowing for stochastic capacity increases and stochastic arrival of a vaccine. Section 6 concludes the paper. All proofs omitted from the body of the paper can be found in the appendix.

2 Setup

Consider a basic susceptible-infectious-recovered (SIR) model with a population whose constant size we normalize to 1. This is a classic model in epidemiology (see, for example, Murray, 2002, p.320), in which the population is divided into three compartments consisting of susceptible individuals, infected individuals and recovered individuals, respectively denoted by $S(t)$, $I(t)$ and $R(t)$ at time t. Note that because of the assumption of a constant population of size 1, for all $t \geq 0$, we have

$$S(t) + I(t) + R(t) = 1.$$

We let $N_1 = S(0)$, $N_2 = I(0)$ and $N_3 = R(0)$ and assume that only two types of transitions are possible: susceptible individuals can become infected and infected individuals recover.[9] (As is standard, "recovered" simply means the individuals are no longer infectious, which occurs either because they gained immunity or died following infection.) Let β denote the rate of infection: the average number of contracts per individual per unit time, multiplied by the probability that the infection is transmitted in a given contact between a susceptible and an infectious individual. We assume that we have a well-mixed or homogeneous population so that $I(t)$ is the fraction of contact occurrences that involve an infectious individual and $S(t)$ is the fraction of contract occurrences that involve a susceptible individual. The *rate of transition* between the susceptible compartment and the infectious compartment is thus given by $\beta I(t)$ and $\beta I(t)S(t)$ is the fraction of the population that is newly infected per unit time.[10]

We denote by $\ell \in [0, 1]$ the severity of the lockdown, with $\ell = 0$ meaning no lockdown and $\ell = 1$ meaning complete lockdown. We assume that ℓ is the choice variable of the policymaker, and with regards to the epidemic, its impact is that it affects the transmission

[9] Provided recovered individuals retain immunity over the course of the policy horizon/pandemic, it seems reasonable to abstract from the possibility that recovered individuals eventually become susceptible.

[10] Epidemiologists frequently refer to this as the "law of mass action" (a term borrowed from chemistry), while economists often refer to this as quadratic search à la Diamond and Maskin (1979).

rate β as follows:

$$\beta(\ell) = \beta_0 + (1 - \ell)\beta_1,$$

where $\beta_0 \geq 0$ is a fixed component of the transmission rate, $\beta_1 > 0$ is a constant, and $\beta(\ell)$ makes the dependence of β on ℓ explicit.

We further assume that individuals recover at rate γ.[11] In SIR models, the parameter $R_0 = \beta/\gamma$ plays an important role in governing the dynamics of an epidemic. Suppose that $N_2 > 0$. Then in this simple model, whenever $R_0 N_1 > 1$ the number of infected individuals will increase from time $t = 0$, resulting in an *epidemic*. If $R_0 N_1 < 1$ then the number of infected individuals will decrease from time $t = 0$ and an epidemic does not occur (alternatively, we can think of the "peak" of the epidemic as occurring at time $t = 0$).

The proportion $\tau \in [0, 1]$ of those who are infected need *treatment*, so that, given $I(t)$ and τ, the number of people requiring treatment at time t is

$$T(t) = \tau I(t).$$

Letting $K > 0$ denote the maximum capacity of the healthcare system to treat COVID-19 patients without reducing the care given to other patients in need, the constraint for managing the epidemic is, for all $t \geq 0$,

$$T(t) \leq K. \tag{1}$$

In the following section, we augment the epidemiology model by an economic production function to analyze tradeoffs involving economics. Specifically, we assume that GDP, denoted Y, is produced using labor L according to the production function $Y = L^\alpha$, where $\alpha \in (0, 1)$ is a parameter that measures labor's productivity, which can be calibrated using labor's income share in national accounts data.[12] Letting $L_0 \geq 0$ denote the amount of labor that is not affected by the lockdown variable ℓ, the amount of labor that is productive given ℓ is

$$L(\ell) = L_0 + (1 - \ell)L_1,$$

where L_1 is the part of the labor that is affected by the lockdown variable ℓ. Thus, $L(0)$ is the pre-lockdown labor supply.[13]

[11] For example, if D is the duration of infection then the rate of recovery is given by $\gamma = 1/D$.

[12] This corresponds to assuming a Cobb-Douglas production function with all input factors other than labor being fixed for the duration of the disease; see, for example, Jehle and Reny (2011).

[13] Assuming a constant population (and hence labor supply) seems like a reasonable approximation for the problem at hand, where a relatively short time period is relevant and the death rate for individuals that participate in the workforce is very low. For example, Verity et al. (2020, Table 1) report dramatic

In Section 4, we will also consider a *heterogeneous agent* version of the model.[14] Specifically, we assume that there is a continuum of types $\theta \in [\underline{\theta}, \overline{\theta}]$ in the population.[15] We denote by F the absolutely continuous distribution of types in the population and by $S(\theta, t)$, $I(\theta, t)$ and $R(\theta, t)$ the density of susceptible, infected and recovered individuals, respectively, of type θ at time t. The density $N(\theta)$ of individuals of type θ thus satisfies, for all $t \geq 0$,

$$N(\theta) = S(\theta, t) + I(\theta, t) + R(\theta, t).$$

We let $N_1(\theta) = S(\theta, 0)$, $N_2(\theta) = I(\theta, 0)$ and $N_3(\theta) = R(\theta, 0)$.

We assume that the type of any given individual is observable and that the policymaker can implement a type-dependent lockdown policy $\ell : [\underline{\theta}, \overline{\theta}] \to [0, 1]$, where $\ell(\theta)$ denotes the severity of the lockdown for individuals of type θ, with $\ell(\theta) = 0$ meaning no lockdown and $\ell(\theta) = 1$ meaning complete lockdown. Similarly to the basic SIR model, we impose a homogeneous mixing assumption. The transmission rate $\beta_\ell(\theta)$ for individuals of type θ under the lockdown policy ℓ is then given by

$$\beta_\ell(\theta) = \beta_0 + (1 - \ell(\theta))\beta_1.$$

For simplicity, we assume that the parameters β_0 and β_1 are independent of θ. Note that in the heterogeneous agent model we make the dependence of the evolution of the epidemic on the lockdown policy ℓ explicit by introducing a subscript ℓ to every variable in the model that depends on ℓ.

We assume that $\tau(\theta) \in [0, 1]$ is the proportion of infected individuals of type θ that require *treatment*. Given a lockdown policy ℓ, the total number of individuals $T_\ell(t)$ that require treatment at time t is given by

$$T_\ell(t) = \int_{\underline{\theta}}^{\overline{\theta}} \tau(\theta) I_\ell(\theta, t)\, dF(\theta).$$

Similarly to the homogeneous agent model, we augment the heterogeneous agent model by an economic production function to analyze tradeoffs involving economics. Specifically, if

differences in age-specific fatality-to-infected ratios. For the group of individuals 60 years old and older, this ratio is 3.28% while for the cohort of individuals in their 20s, it is 0.0309%. Thus, individuals in their 20s are roughly 100 times less likely to die when infected than those past 60. Similarly, Williamson et al. (2020, p.2) report that the "overall cumulative incidence of death 90 days after study start was $< 0.01\%$ in those aged 18-39 years, rising to 0.67% and 0.44% in men and women respectively aged ≥ 80 years"; see also their Figure 2.

[14] See, for example (Murray, 2002, p.320) for an age-structured SIR model.

[15] One can think of the type of an agent as encompassing the relevant observable characteristics of that agent, such as age, gender, occupation and underlying health conditions.

$L_\ell(\theta)$ denotes labor supplied by individuals of type θ under lockdown policy ℓ, then output produced by individuals of type θ is given by

$$Y_\ell(\theta) = A(\theta)L_\ell(\theta),$$

where $A(\theta)$ denotes the productivity of individuals of type θ. Total output Y_ℓ under lockdown policy ℓ is thus given by

$$Y_\ell = \int_{\underline{\theta}}^{\overline{\theta}} A(\theta)L_\ell(\theta)\, dF(\theta).$$

Letting $L_0 \geq 0$ denote the proportion of labor that is not affected by the lockdown variable ℓ, the amount of labor of type θ that is productive given ℓ is

$$L_\ell(\theta) = L_0 + (1 - \ell(\theta))L_1.$$

For simplicity, we again assume that L_0 and L_1 are independent of θ.

3 Homogeneous agent model

We now analyze the dynamics of an epidemic and then derive the minimal lockdown policy ℓ_K necessary to satisfy the constraint K at all times in the model with homogeneous agents.

3.1 Dynamics of an epidemic

In this subsection we treat β as a parameter and study the dynamics of an epidemic in our simple homogeneous SIR model. We also characterize the epidemic peak and derive some useful comparative statics. The dynamics of an epidemic in our simple homogeneous agent SIR model are governed by the following system of non-linear differential equations:

$$\frac{dS(t)}{dt} = -\beta I(t)S(t), \quad \frac{dI(t)}{dt} = \beta I(t)S(t) - \gamma I(t) \quad \text{and} \quad \frac{dR(t)}{dt} = \gamma I(t),$$

with initial conditions $S(0) = N_1$, $I(0) = N_2$ and $R(0) = N_3$. Harko et al. (2014) provided an analytic solution to this system of equations by parameterizing time t by a parameter u. In particular, introducing the integration constants

$$S_0 = N_1 \mathrm{e}^{\beta N_3/\gamma}, \quad u_0 = \mathrm{e}^{-\beta N_3}, \quad \text{and} \quad C_1 = -\beta$$

we have

$$t(u) = \int_{u_0}^{u} \frac{1}{\xi(C_1 - \gamma \log(\xi) + S_0\beta\xi)} d\xi,$$
$$I(u) = 1 - S_0 u + \frac{\gamma \log(u)}{\beta},$$
$$R(u) = -\frac{\gamma \log(u)}{\beta}.$$

Notice that when $u = u_0$ we have $t = 0$ and that u decreases as t increases.[16] We can then back out $S(u)$ using $S(u) + I(u) + R(u) = 1$.

The basic dynamics of an epidemic are as follows. As susceptible individuals become infected and then recover, the stock of susceptible individuals decreases over time and the stock of recovered individuals increases over time. The number of infected individuals initially increases before reaching an epidemic peak and then gradually decreasing. The number of infected individuals stops increasing once the population of susceptible individuals is sufficiently small. An example of a typical epidemic path is shown in Figure 1. Note that unless stated otherwise all figures are drawn for the parameterization $N_1 = 0.999$, $N_2 = 0.001$, $N_3 = 0$, and $\gamma = 1/18$; Figure 1 assumes $\beta = 3.1/18$.[17]

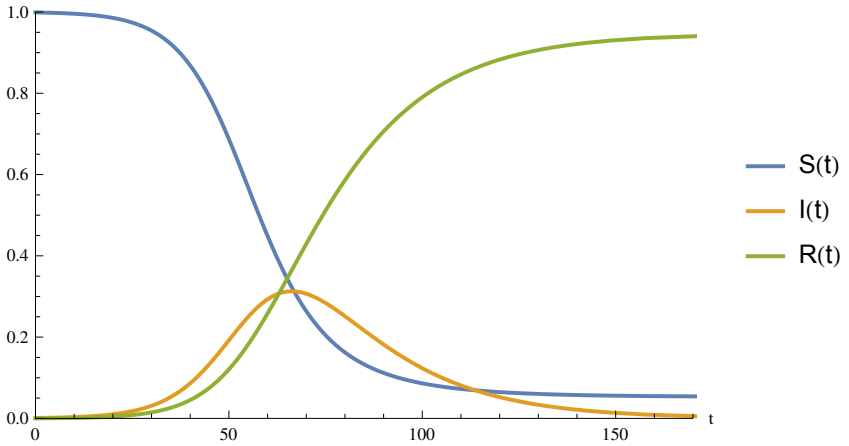

Figure 1: The evolution of a typical epidemic.

Assuming $R_0 N_1 = \beta N_1/\gamma > 1$ (so that the peak of the epidemic does not occur at $t = 0$),

[16] In the limit as $t \to \infty$ we have $u \to -\gamma W\left(-e^{C_1/\gamma}\beta S_0/\gamma\right)/(\beta S_0)$, where W is the product log function.

[17] Here, we assume that 0.1% of the population is initially infected. Following Wang et al. (2020), we take $\gamma = 1/18$, which reflects an average disease duration of 18 days (and so the appropriate interpretation of the time scale t is then also in days). We also set $\beta = \beta_0$ so that $R_0 = 3.1$, which is in line with estimates from Wuhan, China prior to the introduction of strict lockdown measures.

the maximal number of infected individuals $I^*(\beta)$ during the epidemic is characterized by

$$I^*(\beta) = 1 - S_0 u_{\max} + \frac{\gamma \log(u_{\max})}{\beta},$$

where

$$u_{\max} = \frac{\gamma}{\beta S_0}.$$

Notice that we have

$$\frac{d\, I^*(\beta)}{d\, \beta} = -\frac{\gamma \log\left(\frac{\gamma}{\beta N_1}\right)}{\beta^2} > 0, \tag{2}$$

where the inequality follows from the fact that $\log(\gamma/(\beta N_1)) < 0$ since by assumption $\gamma/(\beta N_1) < 1$. Not surprisingly, the peak number of infected individuals, $I^*(\beta)$, increases in the rate of infection β. This means that any policy intervention that decreases β, such as wearing masks, decreases $I^*(\beta)$. Intuitively, and as shown next, this will make it easier to meet a given capacity constraint in the sense that, all else equal, a less severe lockdown is required to satisfy that constraint. Put differently, anyone who dislikes catastrophic health outcomes and lockdowns should be in favour of wearing masks according to this framework.

3.2 Capacity constraints

We now return to the problem faced by our policymaker, as introduced in Section 2, where β is endogenous and depends on the chosen lockdown policy $\ell \in [0, 1]$. Notice that the parameters β_0, β_1, γ and τ, as well as the initial conditions, impose restrictions on the lower feasible bound for K. Specifically, denote by $I_\ell(t)$ the number of infectious at time t given policy $\ell \in [0, 1]$, by

$$I^*_\ell = I^*(\beta(\ell))$$

the maximum number of infected individuals given ℓ, and by

$$T^*_\ell = \tau I^*_\ell$$

the maximum number of people needing treatment per time given policy ℓ. From (2) we have that I^*_ℓ and T^*_ℓ are continuously decreasing in ℓ since $\beta(\ell)$ is decreasing in ℓ. From this and continuity it follows that K is feasible if and only if

$$K \in [T^*_1, T^*_0]. \tag{3}$$

If $K < T_1^*$, then the capacity constraint is so tight that it can never be satisfied at the peak of the epidemic, not even with the most severe lockdown policy. If $K > T_0^*$, then no lockdown is required to satisfy the constraint.

Conversely, for any K satisfying (3) there is a *minimal lockdown policy*, denoted ℓ_K, that satisfies the constraint that the number of individuals requiring treatment at time t never exceeds K. Formally,

$$\ell_K := \min\{\ell \in [0,1] \mid T_\ell^* \leq K\}.$$

Because T_ℓ^* is a decreasing function of ℓ, it follows that ℓ_K is a decreasing function of K. Intuitively, as the capacity constraint K increases, the severity of the required lockdown decreases.

As for policy implications, this means that, all else equal, states or countries with larger capacities can afford less stringent lockdowns. For a given lockdown policy the transmission rate parameter β can also vary substantively between states and countries as the value of this parameter varies with factors such as population density and household composition. Since the maximum number of patients requiring treatment is given by $\tau I^*(\beta)$ and $I^*(\beta)$ is increasing in β (see (2)), it follows that, all else equal, states or countries with larger transmission rates require more stringent lockdowns. Formally, compare two regions, each with capacity K, with transmission rates parameterized by (β_0, β_1) and $(\hat{\beta}_0, \hat{\beta}_1)$ satisfying $\hat{\beta}_i \geq \beta_i$ for $i = 0, 1$, where at least one of these inequalities is strict. Denoting the respective minimal lockdown policies by ℓ_K and $\hat{\ell}_K$, we then have

$$\ell_K < \hat{\ell}_K.$$

In other words, regions with lower transmission rates can afford slacker lockdown policies as is illustrated in Panel (a) of Figure 2. This figure uses the same parameters values as Figure 1 but with $(\beta_0, \beta_1) = (0.5\gamma, 2.6\gamma)$ and $(\hat{\beta}_0, \hat{\beta}_1) = (0.6\gamma, 2.6\gamma)$. As noted at the end of Subsection 3.1, this also means that wearing masks affords slacker lockdown policies in the presence of an airborne, transmittable disease.

The relationship between lockdown and capacity We now look in slightly more detail at the relationship between K and ℓ_K. Panel (b) of Figure 2 plots this relationship, assuming $(\beta_0, \beta_1) = (0.5\gamma, 2.6\gamma)$ and $\tau = 0.11$.[18] As before we set $N_3 = 0$ and $\gamma = 1/18$ but

[18]Following Wang et al. (2020), we use $R_0 = 3.1$ with no lockdown and $R_0 = 0.5$ under the strictest possible lockdown. The first of these R_0 values is an estimate for Wuhan, China prior to any policy interventions by the Chinese government. The second of these R_0 values is an estimate for Wuhan, China under the strictest lockdown measures implemented by the government. We use $\tau = 0.11$, which is consistent with data from New York which showed that around 11% of confirmed coronavirus patients were hospitalized at

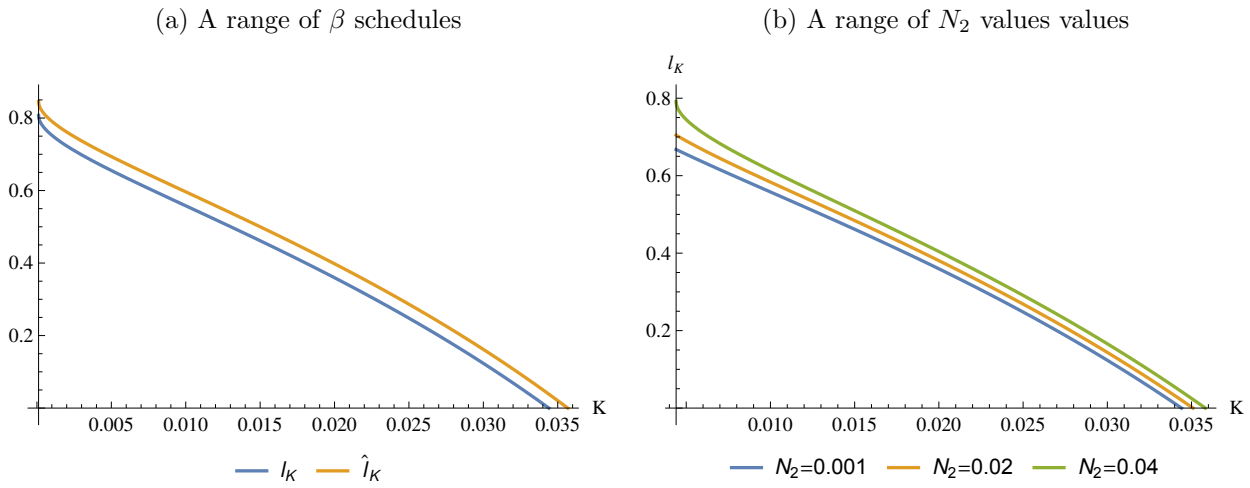

Figure 2: Panel (a) illustrates that a higher schedule of β values necessitates a more severe lockdown for a given K value. Panel (b) illustrates the relationship between ℓ_K and K for a range of N_2 values. As was shown analytically, for a given N_2 value, the severity of the lockdown ℓ_K decreases as the capacity K of the healthcare system increases. This figure also shows that a more severe lockdown is required if a higher proportion N_2 of the population is initially infected.

we now create plots for three different values of N_2: $N_2 = 0.001$ (in which case $N_1 = 0.999$), $N_2 = 0.02$ (in which case $N_1 = 0.98$) and $N_2 = 0.04$ (in which case $N_1 = 0.96$). Panel (b) of Figure 2 shows that the lockdown policy ℓ needed to achieve a given K increases in the proportion of the population that is initially infected. This figure also shows how the proportion of the population that requires treatment at the height of the pandemic, for a given lockdown policy ℓ, increases in the proportion of individuals N_2 that are initially infected. Consequently, for a given cap K, a more severe lockdown is required as N_2 increases. This result highlights the high cost of a delayed policy response.[19]

Figure 3 provides some additional comparative statics showing how the severity of the

the peak in hospitalizations (Feuer, 2020). Note that τ is not the rate of hospitalization (i.e. the proportion of coronavirus patients that are hospitalized at some point over the course of their illness) but rather the proportion of coronavirus patients that are hospitalized at any given point in time. An alternative approach would be to include a separate compartment in the model for hospitalizations, the main advantage being that this would produce a time-lag between the peak in the number of infected individuals and the peak in the number of hospitalized individuals (which is consistent with what we observe in data). We ensure that the binding constraint on the healthcare system is not violated due to this time-lag effect by calibrating the model using the proportion of the population that is hospitalized at the peak in hospitalizations.

[19]For example, when San Francisco issued a shelter-in-place order on March 16, 2020 its number of per capita confirmed coronavirus cases was comparable to that of New York City. By the time New York City was subject to a shelter-in-place order on March 23, 2020 the number of per capita confirmed coronavirus cases was greater than that of San Francisco by roughly an order of magnitude.

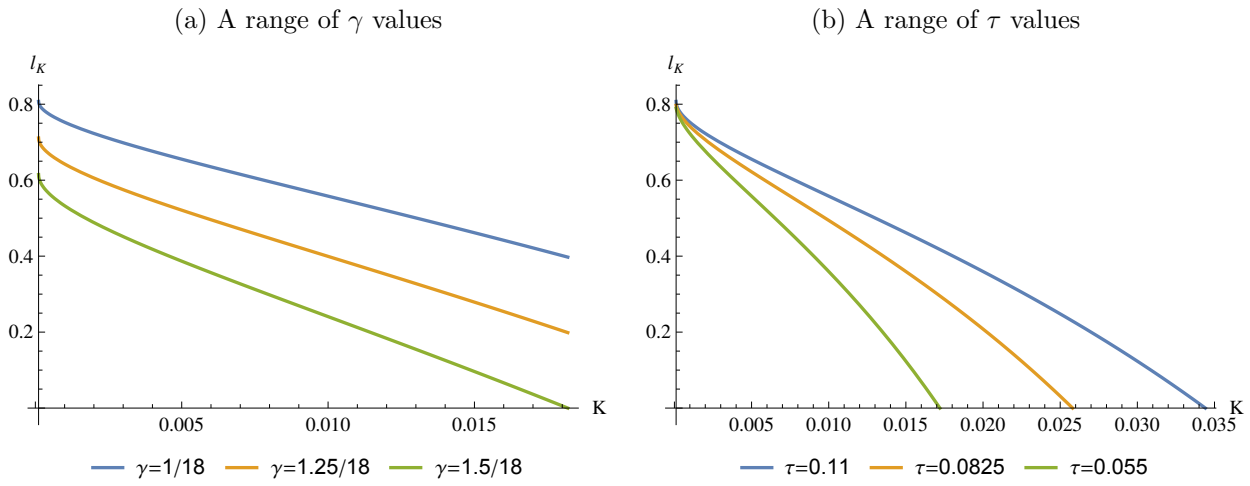

Figure 3: The severity of the required lockdown decreases as γ increases and as τ decreases.

required lockdown decreases as γ increases and as τ decreases. Panel (a) uses the same parameter values as Panel (b) of Figure 2 but with $N_2 = 0.001$ (in which case $N_1 = 0.999$ since we set $N_3 = 0$) and $\gamma = 1/18$, $1.25/18$ and $1.5/18$. Panel (b) also uses the same parameter values as Panel (b) of Figure 2 but with $N_2 = 0.001$ (in which case $N_1 = 0.999$ since we set $N_3 = 0$) and $\tau = 0.11$, 0.0875 and 0.055. One interpretation of these comparative statics is that as superior treatments become available, individuals both require less overall treatment and recover from the disease more quickly and hence a less severe lockdown is required.

Economic impact By mapping the severity of the lockdown to gross domestic product (GDP), one can trace out the relationship between the capacity constraint and economic output. Substituting $L(\ell)$ into the production function yields output

$$Y(\ell) = (L_0 + (1-\ell)L_1)^{\alpha}.$$

It follows that, given the minimal lockdown policy ℓ_K for the constraint K, output, denoted Y_K, is given by

$$Y_K = Y(\ell_K).$$

Because $Y(\ell)$ decreases in ℓ and ℓ_K decreases in K, it follows that

$$\frac{dY_K}{dK} > 0.$$

A plot illustrating how Y_K increases in K for a given set of parameters can be found in Figure 4, which assumes $L_0 = 1/3$, $L_1 = 2/3$ and $\alpha = 1/3$ (and as before $N_2 = 0.001, 0.02, 0.04$, $N_3 = 0$, $\beta(\ell) = 0.5\gamma + 2.6\gamma(1-\ell)$, $\gamma = 1/18$, $\tau = 0.11$).[20] This plot shows that longer delay in the initial policy response—which leads to a higher number of infected individuals in the population prior to any lockdown intervention—results in policymakers facing a more severe economic impact of the pandemic in order to satisfy the binding constraint K.

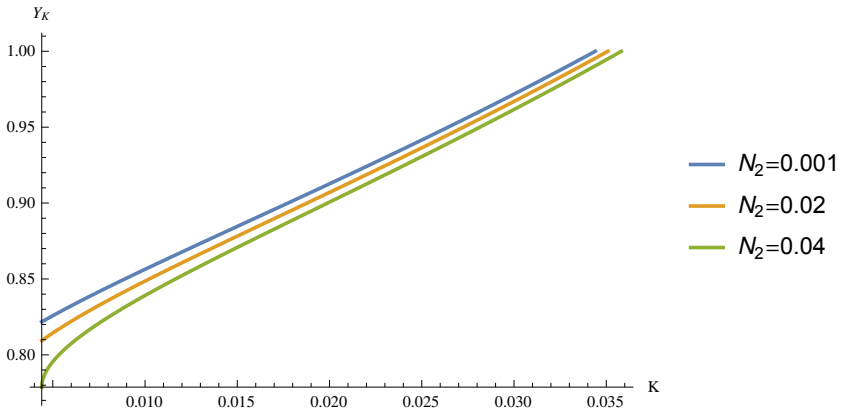

Figure 4: Output Y_K increases in K and decreases in N_2.

3.3 Sensitivity analysis and confidence intervals

The very nature of contagious diseases is that, inherently, their dynamics are non-linear. A large part of the need to use epidemiology models arises from the need to predict future spread and, in our setting, healthcare demand as a policy that only adapts to concurrent data without accounting for states in the future will fail to satisfy the capacity constraints. Of course, as with any predictive model, predictions are subject to uncertainty and errors that can result both from model misspecification and from uncertainty about parameters within the model. We now briefly discuss how the latter can be accounted for within the homogeneous population model.

Up to this point we have treated the parameters of the model as given and known. As just mentioned, in practice, there may be considerable uncertainty and measurement error

[20] We use a standard value of $\alpha = 1/3$ and setting $L_0 = 1/3$ corresponds to a labor force in which one third of workers are essential. While we do not pursue this here, our model also lends itself to the possibility of relating the costs and benefits of a lockdown to the statistical value of life such as DALY (disability-adjusted life year) or QALY (quality-adjusted life year) that are widely used in public health debates. See Hall et al. (2020) for a framework that considers the tradeoff between consumption and COVID-19 deaths. By expressing the value of a life in terms of years of per capita consumption this approach allows those authors to derive an upper bound on the level of consumption a utilitarian society would be willing to forgo in order to avoid COVID-19 deaths.

associated with these parameter values, implying that the predictions of the model are not deterministic. We now discuss how the distribution of the predicted peak of the epidemic can be derived from the, by assumption, known distributions of the uncertain parameters R_0, τ, N_1, N_2 and N_3.

After some tedious algebra, the proportion of the population requiring treatment at the peak of the epidemic is given by

$$T^*(R_0, \tau, N_1, N_3) = \tau \left(1 - N_3 - \frac{\log(R_0) + \log(N_1) + 1}{R_0}\right).$$

The density of T^* is thus

$$f_{T^*}(y) = \int_{\mathbb{R}^5} f_{R_0}(x_1) f_\tau(x_2) f_{N_1}(x_3) f_{N_3}(x_4) \delta(y - T^*(x_1, x_2, x_3, x_4))\, dx_1\, dx_2\, dx_3\, dx_4,$$

where δ denotes the Dirac delta function and f_X denotes the distribution of the random variable X. From here one can construct a confidence interval for the maximum number of individuals requiring treatment at the peak of the epidemic.

For example, suppose that R_0, N_1 and N_3 are known parameters and that $\tau \sim \mathrm{N}(0.11, 0.01)$. That is, τ is normally distributed with mean 0.11 and standard deviation 0.01. Then $T^*(R_0, \tau, N_1, N_3)$ is normally distributed with mean $\mu(R_0, N_1, N_2) = T^*(R_0, 0.11, N_1, N_3)$ and standard deviation

$$\sigma(R_0, N_1, N_2) = 0.01 \left(1 - N_3 - \frac{\log(R_0) + \log(N_1) + 1}{R_0}\right).$$

A 95% confidence interval for the value of T^* is then given by $[\mu - 1.96\sigma, \mu + 1.96\sigma]$. Therefore, if we use $T = \mu(R_0, N_1, N_2) + 1.96\sigma(R_0, N_1, N_2)$ then we can say that with 97.5% confidence, the constraint K will not be violated at the peak of the epidemic. An illustration is provided in Figure 5. This figure uses precisely the same parameters as those shown in Panel (b) of Figure 2 and sets $N_2 = 0.001$ (in which case $N_1 = 0.999$).

An alternative to the approach adopted here would be to perform a conditional worst-case analysis. That is, if one had estimates of moments of a particular parameter distribution (such as its mean and variance), then one could compute confidence intervals with respect to the worst-case distribution.

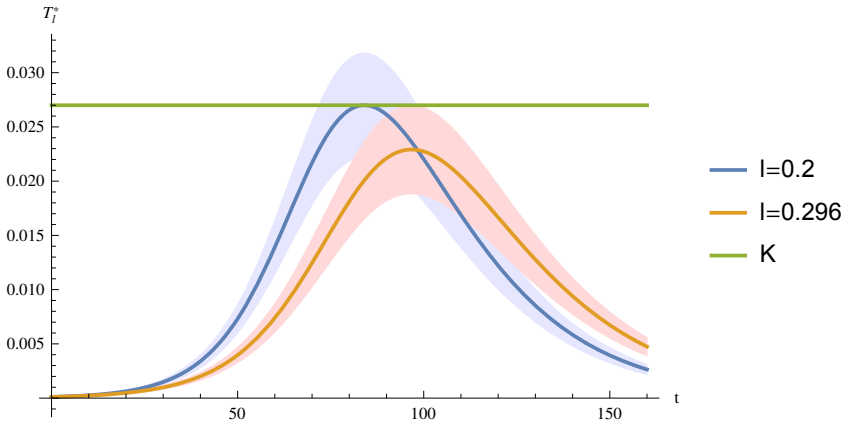

Figure 5: Assume $K = 0.0270$. If τ is deterministic and equal to 0.11, we have $\ell_K = 0.2$. In contrast, if τ is normally distributed with mean 0.11 and a standard deviation of 0.01, then $\ell = 0.296$ is required to satisfy the constraint with a probability of 0.975.

4 Heterogeneous agent model

We now turn our attention to the heterogeneous agent model. Without loss of generality, one can normalize $N(\theta) = 1$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$. Note that this implies that

$$\int_{\underline{\theta}}^{\overline{\theta}} (S(\theta, t) + I(\theta, t) + R(\theta, t))\, dF(\theta) = 1.$$

Under our assumption of homogeneous mixing among all type cohorts, the time t rate of transition between the compartment of susceptible individuals of type θ and the compartment of infected individuals of type θ is

$$\int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y) I_\ell(y, t)\, dF(y).$$

The dynamics of an epidemic in this SIR model are then governed by the following system of non-linear differential equations

$$\begin{aligned}
\frac{dS_\ell(\theta, t)}{dt} &= -\int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y) I_\ell(y, t)\, dF(y) S_\ell(\theta, t), \\
\frac{dI_\ell(\theta, t)}{dt} &= \int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y) I_\ell(y, t)\, dF(y) S_\ell(\theta, t) - \gamma I_\ell(a, t), \\
\frac{dR_\ell(\theta, t)}{dt} &= \gamma I_\ell(\theta, t),
\end{aligned}$$

with initial conditions $S_\ell(\theta,0) = N_1(\theta)$, $I_\ell(\theta,0) = N_2(\theta)$ and $R_\ell(\theta,0) = N_3(\theta)$, where $N_1(\theta) + N_2(\theta) + N_3(\theta) = 1$. Letting $S_\ell(t) = \int_{\underline{\theta}}^{\overline{\theta}} S_\ell(\theta,t)\,d\theta$, $I_\ell(t) = \int_{\underline{a}}^{\overline{\theta}} I_\ell(\theta,t)\,d\theta$ and $R_\ell(t) = \int_{\underline{\theta}}^{\overline{\theta}} R_\ell(\theta,t)\,d\theta$ and integrating this system of differential equations yields

$$\begin{aligned}
\frac{dS_\ell(t)}{dt} &= -\int_{\underline{a}}^{\overline{a}} \beta_\ell(y) I_\ell(y,t)\,dF(y) S_\ell(t), \\
\frac{dI_\ell(t)}{dt} &= \int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y) I_\ell(y,t)\,dF(y) S_\ell(t) - \gamma I_\ell(t), \\
\frac{dR_\ell(t)}{dt} &= \gamma I_\ell(t).
\end{aligned}$$

The social planner then selects the lockdown policy ℓ that maximizes output Y_ℓ subject to the constraint that total hospitalizations not exceed K at any point during the epidemic. That is, for all $t \geq 0$,

$$T_\ell(t) \leq K.$$

To ensure that we have an interesting problem, we assume that this constraint is violated under the lockdown policy $\ell(\theta) \equiv 0$ and is slack under lockdown policy $\ell(\theta) \equiv 1$. Under homogeneous mixing this model reduces to a simple homogeneous agent SIR model with a transmission rate $\beta = \int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y) I_\ell(y,t)\,dF(y)$ that depends on the type-dependent lockdown policy. Exploiting this fact yields the following proposition.

Proposition 1. *Assume $R(\theta,0) = k_0$ and $I(\theta,0)/S(\theta,0) = k_1$, where $k_0, k_1 > 0$ are constants, and for all $\theta \in [\underline{\theta}, \overline{\theta}]$, $\frac{dA(\theta)}{d\theta} < 0$. Then the optimal policy is bang-bang, that is, there is an $\theta^* \in [\underline{\theta}, \overline{\theta}]$ such that*

$$\ell^*(\theta) = \begin{cases} 0, & \theta \leq \theta^* \\ 1, & \theta > \theta^* \end{cases}.$$

Having a bang-bang solution is not only analytically convenient but also useful in practice: even though the planner might want to consider a continuum of lockdown policies, which would pose practical difficulties, in this case it is without loss of generality to only consider minimal and maximal lockdown policies across type cohorts. Figure 6 indicates that similar comparative statics to those illustrated in Figures 2 and 3 hold for the cutoff type θ^* that characterizes the optimal lockdown policy under the heterogeneous agent model.

Interestingly, the heterogeneous agent model induces a non-trivial economic optimization problem that does *not* require taking a stance on how economic activity is traded off against the number of deaths caused by the disease. Indeed, the optimal lockdown policy in this setup maximizes economic output subject to a given capacity constraint. In this sense, the dollars-death tradeoff is not the starting point of the analysis but rather a result of the

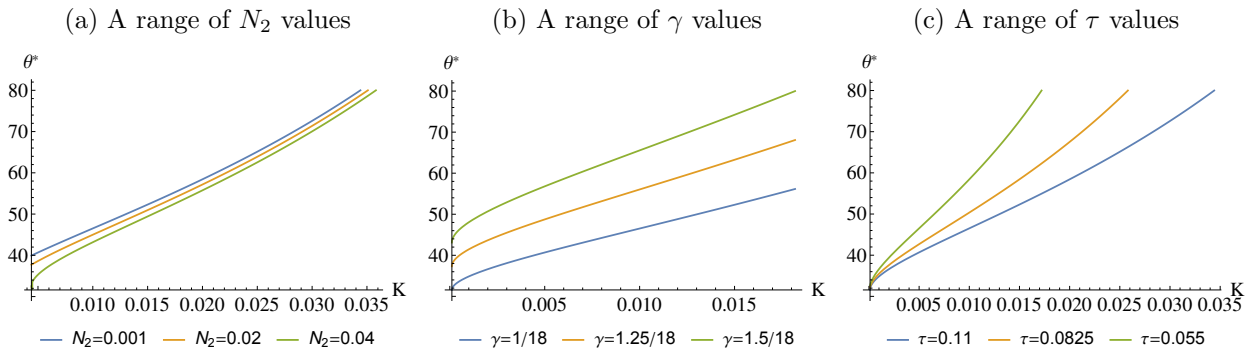

Figure 6: The type cutoff θ^* associated with the optimal bang-bang lockdown policy increases in K and γ and decreases in N_2 and τ. These figures use the same parameter values as Figures 2 and 3 where types are distributed uniformly over the interval $[20, 80]$ and τ is constant across types.

analysis.[21] Of course, many of the specific policy implications derived from our setup need not carry over to richer models but the basic feature that models with heterogeneous agents and a capacity constraint induce an economic optimization problem without specifying the value of life remains valid.

Policy-dependent mixing Another notable feature of Proposition 1 is that it does not include an assumption concerning how τ varies with θ. This is a direct consequence of the homogeneous mixing assumption. However, in practice, one might also expect a lockdown policy to impact how the type cohorts mix. In this case, the structure of the optimal policy will also depend on how hospitalization rates vary across types. We now relax the assumption of homogeneous mixing and consider type-dependent mixing. Motivated by the form of the optimal policy under homogeneous mixing and for purposes of tractability we restrict attention to lockdown policies such that $\ell(\theta) \in \{0, 1\}$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$. We further assume that only types subject to the same lockdown policy mix (and mix in a homogeneous fashion). We then have the following proposition.

Proposition 2. *Assume that for all $\theta \in [\underline{\theta}, \overline{\theta}]$ we have $\frac{dA(\theta)}{d\theta} < 0$, $\frac{d\tau(\theta)}{d\theta} > 0$, $R(\theta, 0) = k_0$ and $I(\theta, 0)/S(\theta, 0) = k_1$, where $k_0, k_1 > 0$ are constants. Then under policy-dependent mixing with $\ell(\theta) \in \{0, 1\}$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$ the optimal policy is monotone. That is, there exists a*

[21] While we require that the capacity constraint is not exceeded at the peak of the epidemic, one could (at the cost of increasing the complexity of the model by incorporating additional dimensions of heterogeneity) augment the model with additional constraints such as requiring that the total number of hospitalizations or deaths do not exceed a given threshold.

$\theta^*_{PD} \in [\underline{\theta}, \overline{\theta}]$ *such that*

$$\ell^*(\theta) = \begin{cases} 0, & \theta \leq \theta^*_{PD} \\ 1, & \theta > \theta^*_{PD} \end{cases}.$$

Moreover, $\theta^*_{PD} > \theta^*$.

Suppose, for illustrative purposes, that agent types correspond to age cohorts and that younger age cohorts are more productive than older age cohorts. Then Proposition 1 shows that under homogeneous mixing the optimal lockdown policy allows younger individuals who are more productive to return to work, while older individuals are subject to a strict lockdown in order to combat the spread of the epidemic. Proposition 2 shows that if older age cohorts are also more vulnerable and more likely to require hospitalization, then the optimal lockdown policy under policy-dependent mixing takes a similar form. Moreover, a bang-bang lockdown policy is more effective, in the sense that a smaller proportion of the population is subjected to a lockdown (and hence economic output is higher) with policy-dependent mixing than without it.

5 Discussion

We now discuss several natural extensions of our modeling approach, with a particular focus on optimal dynamic policies. This allows us to compare our results to those of Alvarez et al. (2020) and Acemoglu et al. (2020).

5.1 Optimal dynamic lockdown policies

We now extend our baseline analysis by allowing the lockdown policy to vary over time and determine the optimal dynamic lockdown policy, subject to the capacity constraint. Since we will be solving this model numerically, for simplicity we now consider a discrete-time version of the model. We let S_t, I_t and R_t denote the respective number of individuals that are susceptible, infected and recovered at time $t \in \mathbb{N}_{\geq 0}$. We again assume that we have a population of a constant size and normalize the size of the population to 1 so that $S_t + I_t + R_t = 1$. Given that $R_t = 1 - S_t - I_t$, we have a model with only two state variables: S_t and I_t. We let $\ell : [0,1]^2 \to [0,1]$ denote the policy function, where $\ell(S_t, I_t)$ specifies the lockdown policy in state (S_t, I_t). All other variables in the dynamic model are defined as they were in the static case. We assume that the policymaker discounts future output according

to the discount factor $\rho \in (0,1)$ and thus solves the following optimization problem:

$$\max_{\ell(\cdot)} \left\{ \sum_{t=0}^{\infty} \rho^t Y(\ell(S_t, I_t)) \right\} \tag{4}$$

$$\text{s.t.} \quad \forall t \in \mathbb{N}, \quad S_{t+1} = S_t - \beta(\ell(S_t, I_t))S_t I_t, \quad I_{t+1} = I_t + \beta(\ell(S_t, I_t))S_t I_t - \gamma I_t, \tag{5}$$

$$\tau I_t \leq K, \tag{6}$$

where the objective function in (4) is the time-discounted sum of output, (5) specifies the laws of motion of the state variables and (6) specifies the capacity constraint. As we did in Section 3, we simply take

$$Y(\ell(S_t, I_t)) = (L_0 + (1 - \ell(S_t, I_t))L_1)^{\alpha}.$$

The Bellman equation associated with this problem thus satisfies

$$V(S_t, I_t) = \max_{\ell(S_t, I_t) \in [0,1]} \{(L_0 + (1 - \ell(S_t, I_t))L_1)^{\alpha} + \rho V(S_{t+1}, I_{t+1})\}$$

$$\text{s.t.} \quad S_{t+1} = S_t - \beta(\ell(S_t, I_t))S_t I_t, \quad I_{t+1} = I_t + \beta(\ell(S_t, I_t))S_t I_t - \gamma I_t, \quad \tau I_{t+1} \leq K,$$

where $V : [0,1]^2 \to \mathbb{R}_{\geq 0}$ denotes the value function.[22] Panels (a) and (b) of Figure 7 provide an illustrative numerical solution for the parameters $\rho = 0.98$, $N_1 = 0.86$, $N_2 = 0.14$, $N_3 = 0$, $\gamma = 1/18$, $(\beta_0, \beta_1) = (0.5\gamma, 2.6\gamma)$, $\tau = 0.11$, $K = 0.027$, $L_0 = 1/3$, $L_1 = 2/3$ and $\alpha = 1/3$.[23]

As is illustrated in Panel (a) of Figure 7, the optimal dynamic lockdown policy imposes a short, sharp lockdown. The policymaker allows the pandemic to progress to the point where the constraint binds before implementing a strict lockdown that prevents the constraint from being violated. The constraint then binds for an extended period while the policymaker swiftly eases the lockdown. Once the constraint becomes slack following the peak of the pandemic we have $\ell = 0$. The qualitative features of the optimal dynamic lockdown policy differs substantively from those derived Alvarez et al. (2020).[24] Their optimal dynamic policies have a "hump-shaped" appearance (see their Figures 1 through 8), with the policymaker gradually easing into and out of the lockdown.

[22] Since this model is solved numerically, one could easily make far more sophisticated assumptions concerning the composition of the workforce at any point in time. For example, one could account for the impact of deaths on the size of the workforce or assume that individuals work at reduced capacity while they are infected. None of these adjustments impact the qualitative features of the model.

[23] Recall that the parameters γ, β_0, β_1 and τ were taken from the epidemiology literature (see footnotes 17 and 18). The values of the parameters α and ρ are ones that are commonly used in the macroeconomics literature. The remaining parameters were chosen to provide an illustrative numerical solution.

[24] As mentioned in the introduction, these authors consider a tradeoff between output costs associated with lockdown and fatalities that occur as a result of the pandemic.

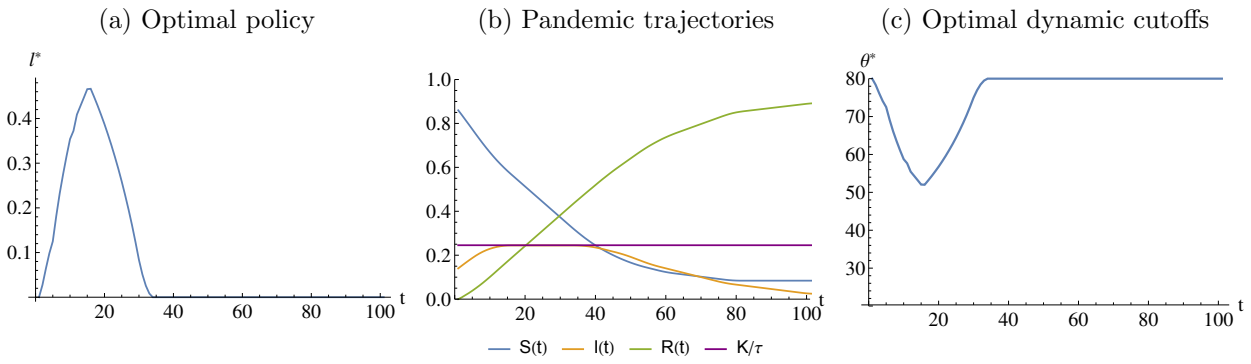

Figure 7: An illustrative solution (Panel (a)) and trajectories (Panel (b)) for the optimal dynamic lockdown policy. Panel (c) displays the optimal cutoff type under a heterogeneous agent model with homogeneous mixing. The cutoff varies over time but at each point in time a bang-bang policy is optimal.

An interesting feature of the dynamically optimal lockdown policy is that once the capacity constraint becomes slack, no future policy interventions are required. Intuitively, the optimal dynamic policy leads to the shortest possible duration of the lockdown by decreasing the population of susceptible individuals as efficiently as possible, subject to the capacity constraint. Consequently, once the capacity constraint is slack we always have $\ell = 0$ under the dynamically optimal policy and no future policy interventions are required. This means that a second wave of infections cannot occur unless (for reasons outside the scope of the model) there is a sufficiently large increase in the population of susceptible individuals. Relative to this optimal dynamic policy derived here, it appears that during the COVID-19 pandemic many US states (such as California) have adopted a policy more akin to a statically optimal policy. Under these policies a longer and less severe lockdown occurs, resulting in an extended period of depressed output. Even after the peak of the pandemic has passed, if a policymaker cancels a statically optimal lockdown too soon, this can result in a large second wave of infections occurring (particularly if the capacity constraint is tight and a large population of susceptible individuals remain after the peak of the pandemic passes). In this sense, statically optimal policies are less robust to future mistakes on the part of policymakers.

In principle, a calibrated version of this dynamic model also allows us to predict total output loss. For example, for the parameters used to construct Panels (a) and (b) in Figure 7, output falls by at most 13% over the course of the pandemic, while the corresponding decrease in the policymaker's objective function (the time-discounted sum of output) is 3.4%. Of course, these predictions are highly sensitive to the choice of parameters and constraint.

The purpose of our framework is to illustrate what we believe to be a fruitful approach to policy rather than produce a precisely calibrated model for which agent-based models à la Akbarpour et al. (2020) are much better suited.[25] That said, with the exception of K and L_0, all the parameter values used for this exercise are, as mentioned, taken from the epidemiology and macroeconomic literature.

The model introduced here can also be extended to account for heterogeneous agents. In particular, if we assume homogeneous mixing among all type cohorts, a bang-bang lockdown will continue to be optimal at every point in time under the same conditions as those stated in Proposition 1. What will change over time is the cutoff θ^* such that types with $\theta > \theta^*$ are subjected to the strictest lockdown possible and types with $\theta < \theta^*$ are not subjected to any lockdown. The optimal policy can then be represented by a function $\theta^*(S_t, I_t)$, which specifies the cutoff type that characterizes the optimal bang-bang lockdown policy in state (S_t, I_t). Panel (c) in Figure 7 illustrates the optimal policy for the same parameters used to construct Panels (a) and (b), with types uniformly distributed over the interval $[20, 80]$ (with τ constant across types). These results are directly applicable to those presented in Section 5.1 of Acemoglu et al. (2020), which considers homogeneous mixing of age cohorts and a "semi-targeted" lockdown.[26] Under a "semi-targeted" lockdown policy, the policymaker can specify one lockdown policy for the oldest age group and another for the young and middle-aged age groups. Since these age groups are fixed, the optimal policies are not bang-bang (see Figures 5.4 and 5.5 of Acemoglu et al. (2020)). In contrast, in our framework, the age groups are endogenous and optimal—an agent is either below or above the threshold implied by the optimal dynamic policy—and the groups vary over time. This shows that when faced with the choice between having (i) a bang-bang policy and time-varying groups or (ii) a rich menu of policies but fixed groups the social planner would prefer (i).

5.2 Arrival of a vaccine

Another important concern for policymakers is the possibility that a vaccine may arrive at some point in the future. We now investigate the implications of this for optimal lockdown policies. Since the arrival of a vaccine is stochastic in nature, it is most natural to consider

[25] This could be a promising avenue for future research as discussed by Akbarpour et al. (2020, pp.22–23) who say: "Importantly, this version of our model does not account for hospital capacity and ICU capacity. This is an important venue for future work, since if our predicted hospitalization and ICU admissions exceed capacity, we expect death rates among symptomatic individuals to increase."

[26] Note, in particular, that the optimality of bang-bang lockdown policies are not specific to our optimization problem which involves a constraint. Such policies would continue to be optimal if we instead considered a tradeoff between the output loss associated with the lockdown and the fatalities and severe health outcomes associated with the pandemic. As emphasized in Section 4, the key assumption that drives the optimality of bang-bang lockdowns is homogeneous mixing of type cohorts.

this in our setting with dynamically optimal policies.

We extend the homogeneous agent model from Section 5.1 by allowing for the stochastic arrival of a vaccine. Specifically, we consider a simple model in which, prior to the arrival of a vaccine, the probability that a vaccine arrives in any given period is $\nu \in (0,1)$. This implies that the arrival period T_ν of the vaccine is geometrically distributed. We assume that when the vaccine arrives, all susceptible individuals are inoculated, which we can represent by moving them to the compartment for recovered individuals.[27] The Bellman equation corresponding to the policymaker's dynamic optimization problem then simply becomes

$$V(S_t, I_t) = \max_{\ell(S_t,I_t)\in[0,1]} \{(L_0 + (1-\ell(S_t,I_t))L_1)^\alpha + \rho((1-\nu)V(S_{t+1},I_{t+1}) + \nu V(0,I_{t+1}))\}$$
$$\text{s.t.} \quad S_{t+1} = S_t - \beta(\ell(S_t,I_t))S_tI_t, \quad I_{t+1} = I_t + \beta(\ell(S_t,I_t))S_tI_t - \gamma I_t, \quad \tau I_{t+1} \leq K.$$

An illustrative numerical solution for the same parameters used in Figure 7 and $\nu = 0.02$ is shown in Figure 8. Naturally, if the vaccine arrives too late (i.e. once the capacity constraint is slack), this does not impact the optimal lockdown policy. However, if the vaccine arrives earlier, this allows the policymaker to immediately ease the lockdown policy, allowing for an earlier economic recovery. The possibility that a vaccine will arrive at some point in the future does not substantively impact the dynamically optimal policy until the arrival actually occurs. Intuitively, this is because the policymaker cannot allow the capacity of the healthcare system to be violated for any realization of the stochastic arrival process. Thus, under a capacity constraint, there are no dynamic complementarities of the form considered in Alvarez et al. (2020), where the possibility that a vaccine will arrive in the future incentivizes the planner to implement a stricter lockdown today.

5.3 Variable healthcare capacity

In all of the models we have considered up to this point, the policymaker knows in advance what the capacity of the healthcare system will be at the peak of the pandemic and chooses its lockdown policy accordingly. However, as we have seen during the recent COVID-19 crisis, many countries expanded the capacity of their healthcare systems over the course of the pandemic. We now analyze the impact of stochastic increases in the capacity of the healthcare system, which is best done in a model with dynamic, rather than static, lockdown policies.

To this end, we now augment the homogeneous agent model from Section 5.1 with the

[27] Assuming that a more extended rollout of the vaccine takes place does not substantively change the qualitative features of the optimal policy or the trajectory of the pandemic.

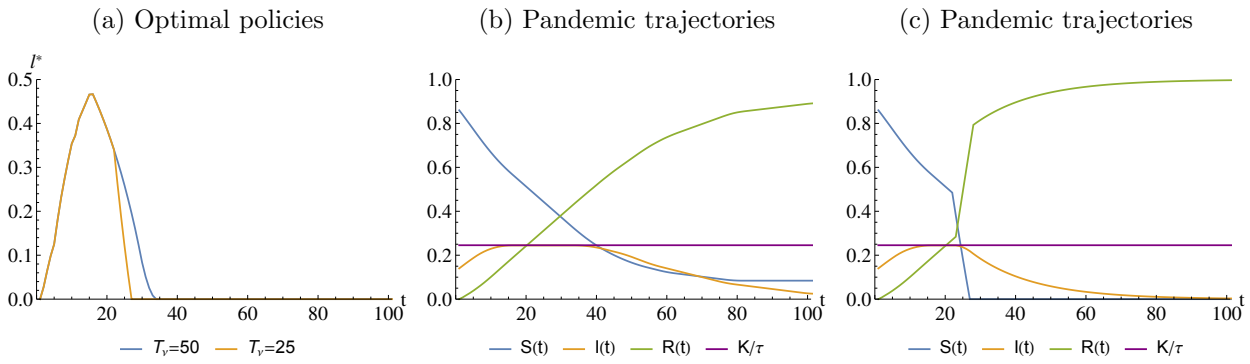

Figure 8: An illustration of the optimal policies (Panel (a)) and pandemic trajectories (Panels (b) and (c)) under the stochastic arrival of a vaccine. Displayed here are two cases involving late arrival of the vaccine (Panel (b), $T_\nu = 50$) and early arrival of the vaccine (Panel (c), $T_\nu = 25$).

stochastic arrival of increased healthcare capacity. Specifically, we consider a model in which the initial capacity of the healthcare system is K_L and at some point the capacity of the healthcare system increases from K_L to K_H, where $K_H > K_L$. In any given period t in which the healthcare capacity is K_L, the probability that the healthcare capacity increases to K_H in period $t+1$ is $\kappa \in (0,1)$. The arrival period T_κ of the increase in healthcare capacity is thus geometrically distributed. Let $V_L : [0,1]^2 \to \mathbb{R}_{\geq 0}$ ($V_H : [0,1]^2 \to \mathbb{R}_{\geq 0}$) denote the value function for states in which the capacity is K_L (K_H). The Bellman equations corresponding to the policymaker's dynamic optimization problem are now

$$V_L(S_t, I_t) = \max_{\ell(S_t,I_t)\in[0,1]} \{(L_0 + (1-\ell(S_t,I_t))L_1)^\alpha + \rho((1-\kappa)V_L(S_{t+1},I_{t+1}) + \kappa V_H(S_{t+1},I_{t+1}))\}$$
$$\text{s.t.} \quad S_{t+1} = S_t - \beta(\ell(S_t,I_t))S_tI_t, \quad I_{t+1} = I_t + \beta(\ell(S_t,I_t))S_tI_t - \gamma I_t, \quad \tau I_{t+1} \leq K_L,$$

and

$$V_H(S_t, I_t) = \max_{\ell(S_t,I_t)\in[0,1]} \{(L_0 + (1-\ell(S_t,I_t))L_1)^\alpha + \rho V_H(S_{t+1},I_{t+1})\}$$
$$\text{s.t.} \quad S_{t+1} = S_t - \beta(\ell(S_t,I_t))S_tI_t, \quad I_{t+1} = I_t + \beta(\ell(S_t,I_t))S_tI_t - \gamma I_t, \quad \tau I_{t+1} \leq K_H.$$

An illustrative numerical solution for the same parameters used in Figure 7 (but with $K_L = 0.02$, $K_H = 0.03$ and $\kappa = 0.02$) is shown in Figure 9. Naturally, if the increase in the capacity of the healthcare system arrives too late (i.e. once the K_L capacity constraint is slack), it has no impact on the course of the pandemic. If the increased capacity of the healthcare system arrives earlier, then this allows the policymaker to temporarily ease the lockdown until the

K_H capacity constraint binds. This ensures that the overall lockdown is both shorter and less strict, which results in higher output from the period in which the increased capacity arrives. Similarly to what we saw with the random arrival of a vaccine, there are no dynamic complementarities and the possibility that the capacity of the healthcare system will increase at some point in the future does not substantively impact the dynamically optimal policy until this increase is actually realized.

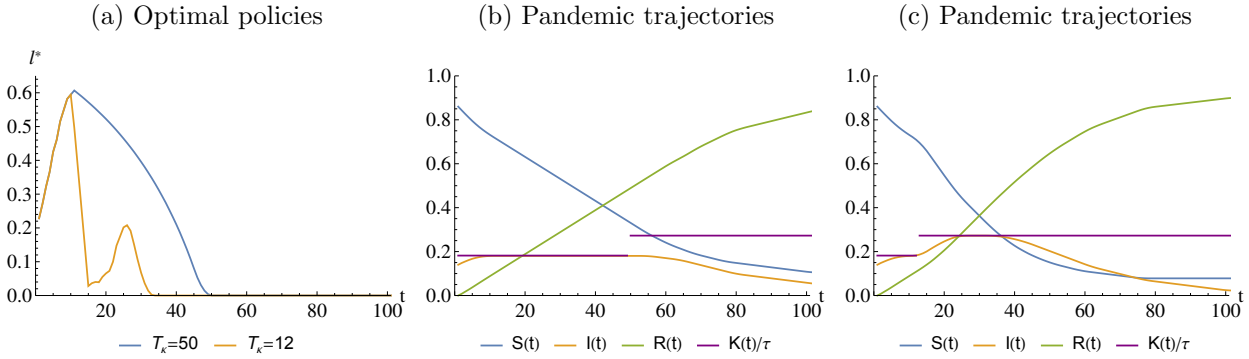

Figure 9: An illustration of the optimal policies (Panel (a)) and pandemic trajectories (Panels (b) and (c)) under stochastic arrival of increased increased healthcare capacity. Displayed here are two cases involving late arrival of the increased capacity (Panel (b), $T_\kappa = 50$) and early arrival of the increased capacity (Panel (c), $T_\kappa = 12$).

6 Conclusions

This time *is* different.[28] The cause of the economic downturn associated with COVID-19 (a pandemic rather than the burst of a financial bubble or any other structural issue with the economy), its scope (universal, hitting all countries more or less within the same quarter) and magnitude (record increases in unemployment filings in the United States) are unprecedented. While there are good reasons to be confident that, informed by the in-depth analyses of past mistakes, the policy response to a severe economic downturn will be better and swifter than at the onset and during the Great Depression, the unparalleled nature of the current shock makes recovery a perilous and winding road. Although policymakers may be ready to act swiftly, the ongoing virulence of the disease may prevent them from so doing. Without widespread immunization, return to normalcy would be difficult if not impossible

[28]This sentence is the title of the New York Times bestseller by Reinhart and Rogoff (2009), where it is used to explain why financial crises occur—because decision makers, in the lead-up to a financial crisis, tend to ignore important precedents. Here and now, however, it seems an accurate description of the current COVID-19 crisis.

even if there were no inertia in rebooting economies that have come to a standstill. We will have to find the path to recovery by learning on the go, and learning quickly.

During the COVID-19 pandemic, arguments have been put forth that policymakers should first take care of the public health aspect of the pandemic and only tackle the economic fallout once the health crisis has been dealt with. Generally speaking, it is not clear what it means to only turn to the economic aspects down the track nor whether the two dimensions can be really separated. Effective COVID-19 vaccines have now been developed, which is a much more comfortable situation than the world was in before. However, as the onslaught of the winter wave in Europe and the United States makes painstakingly clear, there is a time delay between having an effective vaccine and a vaccine becoming effective. Tough months and decisions lie ahead for policymakers in these and many other regions of the world, and catastrophic health outcomes like those New York City or Lombardy experienced in the first half of 2020 remain a lurking threat. In future pandemics, developing an effective vaccine may prove elusive, in which case the health crisis and the economic crisis cannot be separated. Our approach provides a way of formalizing the notion of dealing with the health crisis first—avoid health catastrophes by satisfying the capacity constraints at all times—while minimizing the economic fallout of satisfying these constraints.

Continuum SIR models, such as the ones analyzed here, provide good approximations for large populations. However, for smaller populations or more refined targets—such as ensuring that ICU beds do not run out—this family of models does not necessarily provide a good approximation. For these kinds of applications, models using agent-based simulations are more appropriate tools, and if calibrated to rich micro-level data, provide more reliable estimates of the outcomes of interest. A promising avenue for future research would be to develop agent-based models that account for the healthcare system's capacity constraint. While it is true that, looking backwards, pandemics of the nature of COVID-19 are once-in-a-century events, there is no reason to believe that this continues to be the case going forward, given the growth of world population and globalization of trade and travel. Policies that are based on the backward-looking perspective will not be sustainable when the next pandemic comes around. Having frameworks at hand to guide policy in this contingency will be invaluable.

References

Acemoglu, D., V. Chernozhukov, I. Werning, and M. D. Whinston (2020): "A Multi-Risk SIR Model with Optimally Targeted Lockdowns," NBER Working Paper #27102.

Akbarpour, M., C. Cook, A. Marzuoli, S. Mongey, A. Nagaraj, M. Saccarola, P. Tebaldi, S. Vasserman, and H. Yang (2020): "Socioeconomic Network Heterogeneity and Pandemic Policy Response," NBER Working Paper #27374.

Alvarez, F. E., D. Argente, and F. Lippi (2020): "A Simple Planning Problem for Covid-19 Lockdown," NBER Working Paper #26981.

Atkeson, A. (2020): "What Will Be the Economic Impact of Covid-19 in the US? Rough Estimates of Disease Scenarios," NBER Working Paper #26867.

Brodeur, A., D. Gray, A. Islam, and S. J. Bhuiyan (2020): "A Literature Review of the Economics of COVID-19," IZA Discussion Paper #13411.

Budish, E. (2020): "$R < 1$ as an Economic Constraint: Can We "Expand the Frontier" in the Fight Against Covid-19?" Note.

Diamond, P. A. and E. Maskin (1979): "An Equilibrium Analysis of Search and Breach of Contract, I: Steady States," *Bell Journal of Economics*, 10, 282–316.

Farboodi, M., G. Jarosch, and R. Shimer (2020): "Internal and External Effects of Social Distancing in a Pandemic," NBER Working Paper #27059.

Feuer, A. (2020): "Stark Death Toll, but Cautious Optimism in N.Y. Over Hospitalizations," *New York Times*.

Gilbert, M., M. Dewatripont, E. Muraille, J.-P. Platteau, and M. Goldman (2020): "Preparing for a responsible lockdown exit strategy," *Nature Medicine*, Published Online, April 14.

Hall, R. E., C. I. Jones, and P. J. Klenow (2020): "Trading Off Consumption and COVID-19 Deaths," NBER Working Paper #27340.

Harko, T., F. S. N. Lobo, and M. K. Mak (2014): "Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates," *Applied Mathematics and Computation*, 236, 184–194.

Jehle, G. and P. Reny (2011): *Advanced Microeconomic Theory*, Addison-Wesley, 3rd ed.

Murray, J. D. (2002): *Mathematical Biology: I. An Introduction*, Berlin: Springer, 3rd ed.

Reinhart, C. and K. Rogoff (2009): *This time is different*, Princeton, New Jersey: Princeton University Press.

Stock, J. H. (2020): "Reopening the Coronavirus-Closed Economy," Hutchins Center Working Paper #60.

Verity, R., L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, ..., and N. M. Ferguson (2020): "Estimates of the severity of coronavirus disease 2019: a model-based analysis," *The Lancet*, Published Online, March 30.

Wang, H., Z. Wang, Y. Dong, R. Chang, C. Xu, X. Yu, ..., and Y. Cai (2020): "Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China," *Cell Discovery*, 6.

Williamson, E. J., A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, . Caroline E. Morton, and B. Goldacre (2020): "Open SAFELY: factors associated with COVID-19 death in 17 million patients," *Nature*, 8 July.

Appendix

A Proofs

A.1 Proof of Proposition 1

Before proving Proposition 1 we prove a useful lemma. In particular, in the proof of Proposition 1 we will see that the planner's optimization problem is equivalent to maximizing output Y_ℓ subject to the constraint that the population average transmission rate $\int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(\theta)\, dF(\theta)$ does not exceed $\overline{\beta}(K) > 0$. We therefore start by proving that this equivalent optimization problem has a bang-bang solution. Note that in order to have an interesting problem, we assume that

$$\beta_0 < \overline{\beta}(K) < \beta_0 + \beta_1.$$

This implies that the constraint is violated under the lockdown policy $\ell(\theta) \equiv 0$ and satisfied under the lockdown policy $\ell(\theta) \equiv 1$.

Lemma A.1. *Assume that $\frac{dA(\theta)}{d\theta} < 0$ for every $\theta \in [\underline{\theta}, \overline{\theta}]$ and suppose that the planner maximizes output Y_ℓ subject to the constraint that the population average transmission rate $\int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(\theta)\, dF(\theta)$ does not exceed $\overline{\beta}(K) \in (\beta_0, \beta_0 + \beta_1)$. Then the optimal policy is bang-bang. That is, there exists $\theta^* \in [\underline{\theta}, \overline{\theta}]$ such that*

$$\ell^*(\theta) = \begin{cases} 0, & \theta \leq \theta^* \\ 1, & \theta > \theta^* \end{cases}.$$

Proof. The social planner solves

$$\max_{\ell(\cdot)} \int_{\underline{\theta}}^{\overline{\theta}} A(\theta) L_\ell(\theta)\, dF(\theta) \quad \text{s.t.} \quad \int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(\theta)\, dF(\theta) \leq \overline{\beta}(K).$$

Let $\mu^* \geq 0$ be the solution value of the Lagrange multiplier. By assumption the constraint is binding so $\mu^* > 0$. Substituting $L_\ell(\theta) = L_0 + (1 - \ell(\theta))L_1$, it follows that $\ell^*(\theta)$ solves

$$\max_{\ell(\cdot)} \int \left[A(\theta)(L_0 + (1 - \ell(\theta))L_1) - \mu^*(\beta_0 + (1 - \ell(\theta))\beta_1)\right] dF(\theta) + \mu^* \overline{\beta}(K).$$

We thus have

$$\ell^*(\theta) = \begin{cases} 0, & A(\theta)L_1 \leq \mu^* \beta_1 \\ 1, & A(\theta)L_1 > \mu^* \beta_1 \end{cases}$$

and combining this last expression with the fact that $\frac{dA(\theta)}{d\theta} < 0$ shows that the optimal policy has the desired bang-bang form. □

We are now ready to prove Proposition 1.

Proof. The social planner solves

$$\max_{\ell(\cdot)} \int_{\underline{\theta}}^{\overline{\theta}} A(\theta) L_\ell(\theta)\, dF(\theta) \quad \text{s.t.} \quad \int_{\underline{\theta}}^{\overline{\theta}} \tau(\theta) I_\ell(\theta, t_{\max}(\ell))\, dF(\theta) \leq K,$$

where $t_{\max}(\ell)$ denotes the time at which the maximum number of individuals are hospitalized under the policy ℓ. Next, recall that we introduced the normalization $N(\theta) = 1$ and assume that for all $\theta \in [\underline{\theta}, \overline{\theta}]$ we have the initial conditions $R(\theta, 0) = 0$ and $I(\theta, 0)/S(\theta, 0) = k$, where $k > 0$ is a constant. Then this implies that for all $t \geq 0$ and any $\theta, \theta' \in [\underline{\theta}, \overline{\theta}]$ we have

$$S(\theta, t) = S(\theta', t), \quad I(\theta, t) = I(\theta', t) \quad \text{and} \quad R(\theta, t) = R(\theta', t).$$

This implies that the system of differential equations governing the dynamics of our SIR model can be rewritten

$$\frac{dS(t)}{dt} = -\int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y)\, dF(y) I(t) S(t),$$

$$\frac{dI(t)}{dt} = \int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y)\, dF(y) I(t) S(t) - \gamma I(t),$$

$$\frac{dR(t)}{dt} = \gamma I(t).$$

Thus, the evolution of the epidemic now depends on the lockdown policy ℓ only through the transmission rate $\beta = \int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y)\, dF(y)$. Furthermore, the constraint can be rewritten

$$\tau I_\ell(t_{\max}(\ell)) \leq K,$$

where $\tau = \int_{\underline{\theta}}^{\overline{\theta}} \tau(\theta)\, dF(\theta)$. Thus, it follows from our previous analysis that the constraint is satisfied provided $\int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(y)\, dF(y) \leq \overline{\beta}(K)$, where the upper bound $\overline{\beta}$ on the transmission rate depends on K. Thus, the social planner's problem can be rewritten

$$\max_{\ell(\cdot)} \int_{\underline{\theta}}^{\overline{\theta}} A(\theta) L_\ell(\theta)\, dF(\theta)$$

$$\text{s.t.} \quad \int_{\underline{\theta}}^{\overline{\theta}} \beta_\ell(\theta)\, dF(\theta) \leq \overline{\beta}(K)$$

and it follows from Lemma 1 that the optimal policy has the bang-bang form specified in the proposition statement. □

A.2 Proof of Proposition 2

Proof. The social planner solves

$$\max_{\ell(\cdot)} \int_{\underline{\theta}}^{\overline{\theta}} A(\theta) L_\ell(\theta)\, dF(\theta)$$

$$\text{s.t.} \quad \int_{\underline{\theta}}^{\overline{\theta}} \tau(\theta) I_\ell(\theta, t_{\max}(\ell))\, dF(\theta) \leq K,$$

where $t_{\max}(\ell)$ denotes the time at which the maximum number of individuals are hospitalized under the policy $\ell : [\underline{\theta}, \overline{\theta}] \to \{0, 1\}$. Let $\mu^* \geq 0$ be the solution value of the Lagrange multiplier. By assumption the constraint is binding, so we have $\mu^* > 0$ and hence ℓ^* must solve

$$\max_{\ell(\cdot)} \int_{\underline{\theta}}^{\overline{\theta}} \left(A(\theta)L_\ell(\theta) - \mu^*\tau(\theta)I_\ell(\theta, t_{\max}(\ell))\right) dF(\theta) + \mu^* K.$$

Since we have both $\frac{dA(\theta)}{d\theta} < 0$ and $\frac{d\tau(\theta)}{d\theta} > 0$, it follows that ℓ increases monotonically in θ.

We thus end up with two independent SIR models. For the type cohorts with $\theta \geq \theta^*_{PD}$ we let $S_1(t)$, $I_1(t)$ and $R_1(t)$ denote the respective populations of susceptible, infected and recovered individuals. The dynamics of the epidemic are governed by

$$\frac{dS_1(t)}{dt} = -\beta_0 I_1(t) S_1(t), \quad \frac{dI_1(t)}{dt} = \beta_0 I_1(t) S_1(t) - \gamma I_1(t) \quad \text{and} \quad \frac{dR_1(t)}{dt} = \gamma I_1(t)$$

and for the type cohorts with $\theta < \theta^*_{PD}$ we let $S_0(t)$, $I_0(t)$ and $R_0(t)$ denote the respective populations of susceptible, infected and recovered individuals. The dynamics of the epidemic are governed by

$$\frac{dS_0(t)}{dt} = -\beta_0 I_0(t) S_0(t), \quad \frac{dI_0(t)}{dt} = (\beta_0 + \beta_1) I_0(t) S_0(t) - \gamma I_0(t) \quad \text{and} \quad \frac{dR_0(t)}{dt} = \gamma I_0(t).$$

The path of hospitalized individuals over time is given by

$$T_{\theta^*_{PD}}(t) = \int_{\underline{\theta}}^{\theta^*_{PD}} \tau(\theta)\, dF(\theta) I_0(t) + \int_{\theta^*_{PD}}^{\overline{\theta}} \tau(\theta)\, dF(\theta) I_1(t)$$

and θ^*_{PD} is pinned down by $T_{\theta^*_{PD}}(t) = K$. Under a given bang-bang lockdown policy, by assumption the rate of transmission is lower among higher types under policy-dependent mixing relative to homogeneous mixing. Since higher type cohorts are hospitalized at a higher rate upon becoming infected it immediately follows that $\theta^*_{PD} > \theta^*$. □