Yuting Fu
Wolfson College
University of Oxford

A thesis submitted for the degree of
Master by Research

Trinity 2023

To all people supported me

Acknowledgements

I would like to express my gratitude to my supervisors, Dr.Hanqing Jin and Dr.Ning Wang, for their invaluable guidance, unwavering support, and insightful feedback throughout the entire journey of crafting this master thesis. A special thanks goes to my dedicated collaborator, Dr.Haitao Xiang, whose contributions and teamwork significantly enriched the outcomes of our research group. I extend my appreciation to Dr.Christoph Reisinger, Dr.Justin Sirignano and Dr.Gechun Liang for their thoughtful evaluation and constructive suggestions. I am deeply thankful to LexisNexis and HPCC for their financial support, which made this research possible. Last but certainly not least, I am indebted to my parents for their love, encouragement, and belief in my aspirations.

Abstract

This thesis is a combination of two works using optimal control and imitation learning to address real-world problems in epidemiology and financial markets.

The COVID-19 pandemic has presented unique challenges for policymakers seeking to balance public health and economic impacts. Mathematical models can provide valuable insights to guide lockdown policies. In chapter 1, we propose an extended SIR model incorporating economic decision-making and interactions among susceptible, infected, and recovered populations. An optimal control framework balances infection spread, deaths, and economic activity under lockdown constraints. Experiments highlight trade-offs between short-term recession and long-term benefits. The model provides guidance on lockdown timing, incorporating death costs, and using status information.

In financial markets, historical data provides demonstrations for how expert investors act. Imitation learning offers a data-driven approach to reproduce trading behaviors, without manual reward design needed in reinforcement learning. In chapter 2, we review main imitation learning techniques and applications in finance. Analyses provide error bounds and generalization guarantees.

The final chapter draws conclusions and list open problems for possible future works.

Contents

List of Figures

Chapter 1

Optimal Lockdown Policy for Covid-19: A Modelling Study

1.1 Introduction

For the COVID-19 pandemic, numerous prevention measures have been studied by [21] and [64] in order to control the spread of the virus by the governments. For example, medical measures, such as research on testing, medicine, and vaccine, are accelerated; relatively easy measures, like face masking and social distancing, are also widely accepted and applied. Essentially the most effective prevention of COVID-19 is the lockdown measure which completely ceases the movement of the human being and thus slows down the spread of disease. However, the lockdown measure is incredibly controversial as it imposes a tremendous impact on our society and economy. Hence it might be the most difficult decision to be made by the governments. Especially when and how to impose the lockdown measure is one of the most challenging questions for both politicians and scientists. To address this question, there is a need to develop a mathematical model combining both epidemiology and economics.

Epidemiological models have been widely studied to analyse the dynamics of the pandemic ([41] [66]). However, there is less discussion on how the lockdown policies can influence the economic decisions of people and the spread of disease and how can policymakers make optimal policy in the epidemic. [19] and [18] analyse the government intervention using epidemiological models with exogenous parameters and evaluate the effect of the intervention by simulation results. Some recent papers focus on analysis of optimal policy and policy effect in the framework of the SIR model or its variants. They studied the effect of different measures including fiscal

policy ([15], [17]), testing and quarantine ([52], [6]), intervention policy on multi-aged groups ([9], [1]), social distancing ([37], [16]) and lockdown control ([2], [28], [1]). In previous works, [2] studied the optimal lockdown policy that minimises the value of fatalities and the output costs of the lockdown policy by locking down part of the susceptible and Infectious population, [1] researched the optimal lockdown policies on people of different age groups, and [28] maximise the economic activity level with the burden of the health-care system.

We extend the classic SIR model ([32], [39]) and incorporate an equilibrium framework to study the optimal lockdown policy during the pandemic period. What we innovate from previous works is that they all only took the governments' perspective but did not take people's own reaction to the pandemic and the government policy into consideration, while we adopt the extension to the SIR model from [15] by involving people's economic decision making (consumption and working hours) and embed the SIR model in a simple Cournot equilibrium framework to model people's reaction to each other. Different from [15] that studied the optimal containment policy by controlling the tax rate, we control the level of lockdown, which is more direct and effective for the governments, especially in the early stage of the pandemic. Furthermore, we emphases the cost of death in our model objective of policymakers, which is an important factor in real-world government decision making. Using this method, we can enable the lockdown policy to identify a balance between the impact of the epidemics on the economy and people's health.

The motivation for this work is to address the following questions that the policymakers may face in reality. The main findings are shown as the short answer to these questions.

- What difference does the optimal control make on the economic and health outcome of the epidemic compare to no control? We find that optimal lockdown measures could significantly reduce the deaths and infections caused by the epidemics. Although there is a short-term recession with lockdown control, it has better long-term economic outcomes than doing no control.

- How does the timing of starting and ending affect the optimal lockdown control itself as well as its economic and health consequences? Our results suggest that both the timing of starting and ending the lockdown control policy makes a difference in terms of both the economic and epidemic outcomes. It is best to start the control as early as possible, and it is more important to avoid ending the control too early.

- How does the cost of death affect the lockdown control policy and the outcomes? Whether policymakers regard the deaths as a negative influence on society lead to different results. Regarding deaths as negative results in stricter lockdown control policy which leads to a much better epidemic and slightly worse economic outcomes.

- What if policymakers have additional information on people's health status? Additional information about the health status of people is beneficial, as the optimal separate control on people in different health status will reach much better economic and epidemic outcomes.

1.2 Model

In this section, we first describe the extension to the canonical SIR model. Then analyse the behaviour of susceptible, infectious, and recovered people in regard to their decisions on consumption and working hours under lockdown regulator and formulate the optimal control problem. Finally, we add the cost of death in our model objective. The model primarily considers a Cournot equilibrium framework where agents simultaneously decide their consumption and labor hours, given the prevailing lockdown policy. This approach captures the strategic interactions among different groups (susceptible, infected, and recovered) and the government, leading to a dynamic equilibrium that evolves over time with the pandemic situation.

1.2.1 Extension of SIR

As shown in the classic SIR model ([39] [33]), we classify people into three categories according to [32]:

- Infectious (I) are those who are tested positive to the virus;

- Recovered (R) are those who have been tested positive to the virus and now recovered;

- Susceptible (S) are those who have not been tested positive to the virus.

We assume that all susceptible people are subjects to be infected with some possibility in direct contact with infectious people, and infectious people will recover with a constant probability of π_r or become dead with another constant probability π_d. Our extension is on the infection. All infection happens via direct contact between

susceptible people and infected ones into three types of activities: purchasing and/or consumption of goods and services, working with other people, and other daily activities. A Lockdown policy can be applied to control the working contact, hence change the income flow, which indirectly imposes constraints on the purchasing and consumption.

We use the following equation (1—5) to describe our extended SIR model for the transition among Susceptible, Infected, Recover, and the death outcome.

$$\begin{aligned} T_t &= \pi_{s1}\left(S_t C_t^s\right)\left(I_t C_t^i\right) + \pi_{s2}\left(S_t N_t^s\right)\left(I_t N_t^i\right) + \pi_{s3} S_t I_t, && (1.1)\\ S_{t+1} &= S_t - T_t, && (1.2)\\ I_{t+1} &= I_t + T_t - (\pi_r + \pi_d) I_t, && (1.3)\\ R_{t+1} &= R_t + \pi_r I_t, && (1.4)\\ D_{t+1} &= D_t + \pi_d I_t. && (1.5) \end{aligned}$$

In this system of equations, S_t, I_t, R_t and D_t represents the number of people in categories of Susceptible, Infectious, Recovery and Death respectively at time t. We use (C_t^s, N_t^s) to model the (average) consumption behaviour and working hours of susceptible people, (C_t^i, N_t^i) to model the (average) consumption behaviour and working hours of infectious people, and (C_t^r, N_t^r) to model the (average) consumption behaviour and working hours of a recovered people. T_t in equation (1.1)is the number of newly infectious people in the time period t to $t+1$ and the three terms in the right-hand side of this equation are used to describe the infection by the three different contact between susceptible people and infectious people via consumption, working, and other types of contact.

We use several constant parameters to describe the transition rate between different categories. π_{s1} reflects the transition rate for a susceptible people get infected by infectious people from direct contact via purchasing/consuming. Similarly, π_{s2} reflects the transition rate from direct contact via working, and π_{s3} reflects the transition rate from other contacts.

Denote $\Delta Y_t = Y_{t+1} - Y_t$ for $Y = S, I, R$, then the dynamics of the SIR model is

$$\begin{aligned} \Delta S_t &= -T_t, \\ \Delta I_t &= T_t - (\pi_r + \pi_d) I_t \\ \Delta R_t &= \pi_d I_t. \end{aligned}$$

We use vectors and matrices to simplify our presentation. Denote $X_t = (S_t, I_t, R_t)^\top$, $C_t = (C_t^s, C_t^i, C_t^r)^\top, n_t = (n_t^s, n_t^i, n_t^r)^\top$, and for any $x = (x_1, x_2, x_3)^\top, c = (c_1, c_2, c_3)^\top, n =$

$(n_1, n_2, n_3)^\top$, define

$$T(x, c, n) = x_1 x_2 \left(\pi_{s1} c_1 c_2 + \pi_{s2} n_1 n_2 + \pi_{s3}\right) \tag{1.6}$$

$$F(x, c, n) = \left(-T(x, c, n), T(x, c, n) - (\pi_r + \pi_d) x_2, \pi_d x_2\right)^\top \tag{1.7}$$

then the system can be described as

$$\Delta X_t = F(X_t, C_t, n_t). \tag{1.8}$$

Remark: In the model, we assume the probability of getting infectious by consumption is linearly depend on the consumption time of infectious and susceptible people, which is a over simplification. Incorporating an infection risk from consumption activities such as dining out or shopping into the model is mathematically feasible and could add a layer of realism. This extension would involve modifying the infection dynamics in the SIR model to account for increased exposure risk associated with these activities rather than the consumption time as in the model. The practical challenge lies in accurately quantifying this risk and its impact on the infection rate. Nevertheless, including such a risk could yield more nuanced insights into the interplay between economic activities and the spread of infection, potentially leading to more effective policy recommendations.

1.2.2 Behaviour of individuals in different categories

We study the rational behaviour of all people who maximise their own welfare by choosing proper consumption and working hours like in a normal time, i.e., the virus does not change people's rationality and preference. Also, we use the following utility function to model the utility from consumption and working of an individual,

$$u(c, n) = \ln c - \frac{\theta}{2} n^2 \tag{1.9}$$

where c is the consumption, and n is the working hours. In this utility, the first term measures the utility from consumption, the second term measures the disutility from working, and θ is the weight between this two terms. The form of the utility is primarily for mathematical convenience. The quadratic term in labor hours, n, incorporates increasing marginal disutility of labor, reflecting realistic labor dynamics. This functional form, despite its initial lack of intuitive appeal, provides a balance between analytical tractability and the capacity to realistically model agent decisions under different economic scenarios. It also gives an estimation of θ, as will be clarify shortly. Denote by A the average wage per hour of a person, hence the labor income

of an individual, with working hour n is $A * n$, which will be the upper bound of the consumption, i.e. $An \geq c$.

Denote by n_0 the full working hours in a unit time before the spread of the virus, which is officially guided by the government. It is natural that n_0 is set optimally for the society, and the optimality brings some information of the parameter θ_0. If a person follows the full working hours n_0 optimally, then her labor income will be An_0. Since the utility function is strictly increasing in the consumption, all labor income should be consumed up, hence the optimal consumption c_0 should be $c_0 = An_0$. Then by the optimality of n_0, we have $\frac{\partial u(c_0,n_0)}{\partial n} = \frac{1}{n_0} - \theta n_0 = 0$, by which we will choose θ by

$$\theta = 1/n_0^2.$$

The total utility of a flow of consumption and working hours $\{(c_\tau, n_\tau)\}_{\tau=t,\cdots,T}$ is defined by

$$U(c_\cdot, n_\cdot) = \sum_{t=\tau}^{T} \beta^\tau u\left(c_\tau, n_\tau\right) \tag{1.10}$$

To contain the spreading of the virus, governments need to apply a lockdown policy to reduce direct contacts between people, which will impose stricter constraints on their behaviour. In this paper, we study the lockdown policy by a constraint on the ratio $L \in [0, 1]$ of the working hour in the full working capacity, i.e., given the full working hours n_0, the maximal working hour cannot exceed $n_0 * L$. We suppose the government cannot easily identify individuals into their categories so that the lockdown constraint on the working hours is the same for all people. We formulate the decision making problem for each category with a given lockdown policy $L_\cdot$, and then study the lockdown policy-making problem for the government.

1.2.2.1 Optimal decision of recovered people

Suppose the lockdown measure $L_t \in [0, 1]$ is given for any time t.

A recovered individual aims at maximising his total utility

$$J^r(c_\cdot^r, n_\cdot^r; t) = \sum_{\tau=t}^{T} \beta^{\tau-t} u(c_\tau^r, n_\tau^r) \tag{1.11}$$

with the constraint $c_\tau^r \leq An_\tau^r$ and $n_\tau^r \leq n_0 L_\tau$.

Theorem 1. *At time t with state X_t and the lockdown policy $\{L_\tau : \tau \in [t, T]\}$, the optimal (c^r, n^r) is*

$$c_\tau^{r*} = An_0 L_\tau, n_\tau^{r*} = n_0 L_\tau, \quad \tau = t, \cdots, T.$$

Proof. Since $\frac{\partial J^r(c^r,n^r;t)}{\partial c^r_\tau} = \beta^{\tau-t}\frac{1}{c^r_\tau} > 0$, we have $c^{r*}_\tau = An^r_\tau \ \forall\tau = t,...,T$. Denote $f(c^r,n^r,\lambda^r_n;t) = J^r(c^r,n^r;t) + \sum_{\tau=t}^T \lambda^r_{n\tau}(n_0L_\tau - n^r_\tau)$. Then by KKT condition, $\forall\tau = t,...,T$,

$$\frac{\partial f(c^{r*},n^r,\lambda^r_n;t)}{\partial n^r_\tau} = 0 \Rightarrow \beta^{\tau-t}(-\theta n^r_\tau + \frac{1}{n^r_\tau}) - \lambda^r_{n\tau} = 0$$
$$\lambda^r_{n\tau}(n_0L_\tau - n^r_\tau) = 0, \lambda^r_{n\tau} \geq 0$$

Since $n_0^2\theta = 1$,

$$\lambda^r_{n\tau} = \beta^{\tau-t}(-\theta n^r_\tau + \frac{1}{n^r_\tau}) > \beta^{\tau-t}(-\theta n_0 + \frac{1}{n_0}) = 0$$

Thus $n^{r*}_\tau = n_0L_\tau, c^{r*}_\tau = An_0L_\tau \forall\tau = t,...,T$ □

Notice that the behaviour of recovered people $(c^r_\cdot, n^r_\cdot)$ plays no role in the spread of the virus, hence the behaviour of recovered people will not affect people in other categories. This is why we start to form this easy-to-handle category.

1.2.2.2 Optimal behaviour of infectious people

Similar to the case of recovered people, infectious people also need to choose their optimal consumption and working hours $\{(c^s_t, n^s_t)\}_{t=0,1,\cdots,T}$ to maximise their total utility from consumption and working hour, subject to the constraint that the consumption c^s_t cannot exceed the labour income for the working hour n_t, and n_t must be no more than the lockdown policy $n_0 * L_t$.

The labor income of an infectious people is different from other categories. Because they are infected, their health condition is usually worse than other people. So we introduce a constant ϕ to discount their working efficiency, and the labor income from n_t working hour will be $A * \phi * n$. Furthermore, since an infectious people will have a constant probability π_r to recover and suffer a possibility π_d of death, we need to calculate the distribution over all categories at a future time. For an infectious people at time t, he has the probability π_r to recover in the next unit time, π_d to die, and the rest probability $1-\pi_r-\pi_d$ to stay in the infected category. By this evolution, we can get the conditional probabilities for his health state at a future time $\tau > t$. Denote by $p^{i,i}(t,\tau)$ the probability for him being still infected, $p^{i,r}(t,\tau)$ the probability being recovered, and $p^{i,d}(t,\tau)$ the probability of being dead. Then we can deduce that

$$p^{i,i}(t,\tau) = (1-\pi_r-\pi_d)^{\tau-t}, \tag{1.12}$$
$$p^{i,r}(t,\tau) = \pi_r\frac{(1-(1-\pi_r-\pi_d)^{\tau-t})}{\pi_r+\pi_d}, \tag{1.13}$$
$$p^{i,d}(t,\tau) = \pi_d\frac{(1-(1-\pi_r-\pi_d)^{\tau-t})}{\pi_r+\pi_d}. \tag{1.14}$$

If he recovered, he should behave optimally as a recovered people, while if death has happened unfortunately, we cease the accumulation of any utility. So, for a given flow $(c^i_\cdot, n^i_\cdot)$ of consumption and working hours taken by the infectious people from time t, the accumulated utility he can get will be

$$J^i(c^i_\cdot, n^i_\cdot; t) = \sum_{\tau=t}^{T} \beta^{\tau-t} \left[p^{i,i}(t,\tau) u(c^i_\tau, n^i_\tau) - p^{i,r}(t,\tau) u(c^{r*}_\tau, n^{r*}_\tau) \right]. \tag{1.15}$$

where (c^{r*}, n^{r*}) is the optimal behaviour of a recovered people determined in the previous case, and $p^{i,i}, p^{i,r}$ are as defined in equation (1.12, 1.13).

Theorem 2. *Given the lockdown policy $L_\cdot$, the optimal (c^r, n^r) is*

$$c^{i*}_\tau = A\phi n_0 L_\tau, n^{i*}_\tau = n_0 L_\tau, \quad \tau = t, \cdots, T. \tag{1.16}$$

Proof. Notice $\frac{\partial J^i(c^i, n^i; t)}{\partial c^i_\tau} = (\beta(1-\pi_r-\pi_d))^{\tau-t} \frac{1}{c^i_\tau} > 0$, we have $c^{i*}_\tau = A\phi n^i_\tau \ \forall \tau = t, ..., T$. Denote $f(c^i, n^i, \lambda^i_n; t) = J^i(c^i, n^i; t) + \sum_{\tau=t}^{T} \lambda^i_{n\tau}(n_0 L_\tau - n^i_\tau)$. Then by KKT condition, $\forall \tau = t, ..., T$,

$$\frac{\partial f(c^{i*}, n^i, \lambda^i; t)}{\partial n^i_\tau} = 0 \Rightarrow (\beta(1-\pi_r-\pi_d))^{\tau-t}(-\theta n^i_\tau + \frac{1}{n^i_\tau}) - \lambda^i_{n\tau} = 0$$

$$\lambda^i_{n\tau}(n_0 L_\tau - n^i_\tau) = 0, \lambda^i_{n\tau} \geq 0$$

Since $n_0^2 \theta = 1$,

$$\lambda^i_{n\tau} = (\beta(1-\pi_r-\pi_d))^{\tau-t}(-\theta n^i_\tau + \frac{1}{n^i_\tau}) > (\beta(1-\pi_r-\pi_d))^{\tau-t}(-\theta n_0 + \frac{1}{n_0}) = 0$$

Thus $n^{i*}_\tau = n_0 L_\tau, c^{i*}_\tau = A n_0 \phi L_\tau \forall \tau = t, ..., T$

□

Different from the recovered case, the behaviour of an infectious people (c^i, n^i) is involved in our extend SIR model for the spreading of the virus, hence they will make the decision problem for susceptible people much harder.

Remark: The optimal decisions for infectious and recovered individuals are considered trivial in this model due to the specific assumptions and structure employed. For recovered individuals, their behavior does not impact the spread of the virus, thus simplifying their decision-making process to a standard utility maximization problem. For infectious individuals, the assumption is that their decision-making process is significantly constrained by their health status and the lockdown policy. While this simplification aids in analytical clarity, it may overlook the nuances of real-life

decision-making in these groups. In complex real-world applications, it would be worth considering more complex models that account for varied behaviors and preferences among infectious and recovered individuals, although this would increase the model's complexity and computational demands.

1.2.2.3 Behaviour of susceptible people

The decision planning for a susceptible people from time t is much more complicated if we consider the possibilities for this people to turn into infectious, recovered, and death at different future time spots. We avoid the complexity by taking advantage of the optimal value function for an infected, and model the objective function of a susceptible people recursively.

As for the previous two categories, we start from time t and pick up a susceptible person. Denote the state of the SIR model at the starting time as X_t, and the lockdown policy is fully given as $L_\cdot$).

Suppose he will follow a given flow of consumption and working hours $(c^s_\tau, n^s_\tau)_{\tau=t,t+1,\cdots,T}$ before being infected, and then follow the optimal behaviour after been infected, i.e., his consumption and working hours after infected will switch to the optimal control for an infected person from the infection time. We denote his objective value as

$$\begin{aligned} J^s(c^s_\cdot, n^s_\cdot; t, X_t, L_\cdot) &= u(c^s_t, n^s_t) + \beta\tau_t J^{i*}(t+1, L_\cdot) \\ &\quad +\beta(1-\tau_t) J^s(c^s_\cdot, n^s_\cdot; t+1, X_{t+1}, L_\cdot), \qquad (1.17) \\ J^s(c^s_\cdot, n^s_\cdot; T, X_T, L_\cdot) &= u(c^s_T, n^s_T), \qquad (1.18) \end{aligned}$$

where $\tau_t = \pi_{s1} n_0 A\phi I_t L_t c^s_t + \pi_{s2} n_0 I_t L_t n^s_t + \pi_{s3} I_t$ is the probability of a susceptible person to be infected in the next unit time, $J^{i*}(t+1, L_\cdot)$ is the optimal objective value achievable for an infected person starting from time $t+1$, and X_{t+1} is the SIR state at time $t+1$ resulted by people's behaviour $(c^s_t, n^s_t, c^{i*}_t, n^{i*}_t, c^{r*}_t, n^{r*}_t)$ and the time t state X_t.

Now it is natural that we aim at maximising the objective $J^s(c^s_\cdot, n^s_\cdot; t, X_t, L_\cdot)$ over feasible control flow $(c^s_\cdot, n^s_\cdot)$), ,i.e., the optimal behaviour of a susceptible people will be the solution for the optimisation

$$\begin{array}{ll} \max & J^s(c^s_\cdot, n^s_\cdot; t, X_t, L_\cdot) \\ s.t. & c^s_\tau \le A n^s_\tau, \quad n^s_\tau \le n_0 L_\tau, \qquad \forall \tau \in \{t, t+1, \cdots, T\}. \end{array} \qquad (1.19)$$

Theorem 3. *At time t with state X_t and the lockdown policy $\{L_\tau : \tau \in [t, T]\}$, the optimal (c^s, n^s) is*

$$c^{s*}_\tau = A n^{s*}_\tau, \quad \tau = t, \cdots, T. \qquad (1.20)$$

Proof. We fix the lockdown policy $L.$ and omit it when no confusion will arise.

Denote the value function as $V(t, X_t) = J^s(c_\cdot^{s*}, n_\cdot^{s*}; t, X_t, L.)$. According to the dynamic programming principle, we know V must satisfy

$$\begin{aligned} V(t, X_t) &= \max_{c_t^s \le An_t^s,\, n_t^s \le n_0 L_t} \left[u(c_t^s, n_t^s) + \beta\tau_t J^{i*}(t+1, L.) + \beta(1-\tau_t)V(t+1, X_{t+1})\right] \\ &= u(c_t^{s*}, n_t^{s*}) + \beta\tau_t^* J^{i*}(t+1, L.) + \beta(1-\tau_t^*)V(t+1, X_{t+1}^*), \end{aligned}$$

where τ_t^* and X_t^* are the corresponding infection probability and time $t+1$ state of the SIR model.

If $c_t^{s^*} < An_t^{s^*}$, then, due to the strictly increasing properties of τ_t in both c^s and n^s, we can easily find a value $m \in (c_t^{s^*}, An_t^{s^*})$, and construct another control $c_t^s = m$ and $n_t^s = m/A$, such that the corresponding τ_t will be the same as τ_t^*, hence X_{t+1} will also be the same as X_t^*. But since $c_t^s > c_t^{s*}$ and $n_t^s < n_t^{s*}$, we have $u(c_t^s, n_t^s) > u(c_t^{s*}, n_t^{s*})$, which contradicts the optimality of (c^{s^*}, n^{s^*}) in the dynamic programming principle. □

1.2.3 Optimal Control of the Policymaker

With the optimal behaviour in each category under a given lockdown policy $L.$, we can easily formulate the optimal policy-making problem into an optimal control problem.

Suppose we start the lockdown problem from some time t_0 with the contamination state X_{t_0} being given by $S_{t_0} = s, I_{t_0} = i$ and $R_{t_0} = r$, then the optimal lockdown policy should be the optimal control problem

$$\max_{L.} \quad J^0(L.; t, X_t) = \sum_{t=t_0}^{T} \beta^{t-t_0} \left[S_t u(c_t^{s*}, n_t^{s*}) + I_t u(c_t^{i*}, n_t^{i*}) + R_t u(c_t^{r*}, n_t^{r*})\right], \tag{1.21}$$

where (c_t^{ca*}, n_t^{ca*}) are the optimal consumption and working hours for people in category ca (ca can be s, i or r), which are all determined in previous optimisation problems.

In previous objective J^0, we remove all cases of death. In reality, since death of disease causes has a strong negative impact to a household as well as to the society, regulators should not ignore any death case. We include the strong impact of death cases by introduce a penalty term into the objective

$$J^\lambda(L.; t, X_t) = \sum_{\tau=t}^{T} \beta^{\tau-t} \left[S_\tau u(c_\tau^{s*}, n_\tau^{s*}) + I_\tau u(c_\tau^{i*}, n_\tau^{i*}) + R_\tau u(c_\tau^{r*}, n_\tau^{r*}) - \lambda D_\tau u(c_\tau^{r*}, n_\tau^{r*})\right], \tag{1.22}$$

In this new objective, we measure the the cost of a death by a multiple of the optimal utility for a recovered people, and the multiple $\lambda > 0$ can be viewed as the severity of death in the government's view. When $\lambda = 0$, J^λ reduces to our previous objective J^0.

With this new objective, the problem for a regulator is to solve

$$\begin{aligned} \max_{L_\cdot} \quad & J^\lambda(L_\cdot; t, X_t), \\ s.t. \quad & L_t \in [0,1] \ \ \forall t \in [0,T]. \end{aligned} \tag{1.23}$$

1.2.4 Solving Scheme

In Problem (1.23), or its reduced version (1.21), the optimal decisions of individuals in all three categories are involved. The optimal decisions for infectious and recovered individuals are considered trivial in this model due to the assumptions and structure employed, which leaves us to tackle the optimal decision problem (1.19) for susceptible people before the Problem (1.23).

We start our solving scheme by tackling the Problem (1.19) with a given lockdown policy $L_\cdot$. Because of the lockdown constraint, it is almost hopeless for us to get an explicit solution. We solve this optimal control problem numerically in the same was as in [15]. In this approach, the optimal control at each time step is regarded as the static optimisation with two constraints from the consumption budget and the lockdown policy on the working hours, and solutions are obtained by solving the corresponding KKT condition[1].

With the optimal control $(c_\cdot^{s*}, n^{s^*})$ as functions of the lockdown policy $L_\cdot$, we deal with the optimal control problem (1.23) as an optimisation over the high dimension space $[0,1]^T$ by the gradient-based interior-point method used in the Matlab function `fmincon`. Although we have no theoretical proof on the convergence of our scheme, our numerical results show the convergence of our scheme.

Parts of our code in our scheme are from [15].

1.3 Model Parameters

In this section, we study how to estimate those parameters in our model from real data, and apply it in an example with COVID-19 data in the UK to get the numerical

[1]In fact, when we use the numerical scheme proposed in [15] to our problem, the derivative used in the KKT condition is not correct due to the absence of a complicated term from the term in equation (1.17). We decide to ignore this absence due to the following two reasons: (1) if we recover this complicated term, the calculation will be extremely complicated; (2) from real data in the COVID-19 pandemic, we know the coefficient in the third term $\beta\tau_t$) is very close to 0, which is also observed in our numerical results.

results for optimal lockdown control.

In our model, we have quite a lot of parameters, and some of them are well-estimated and available from different sources. Let us start from easily accessible ones.

For the extended SIR model, without loss of generality, we standardise the total population to $N = 1$, which makes S_t, I_t, R_t and D_t be the proportions of the population of each category in the total population.

The unit of a time step is not an essential parameter, we can simply count the time by weeks.

π_r and π_d in the extended SIR model can be easily estimated from historical data, which have been done in several data sources [2]. In our example, we will use the estimation from [15].

π_{s1}, π_{s2} and π_{s3} are complicated to estimate, and we defer the discussion to after all easy ones.

For the characterisation of the decision making for individuals, we still need parameters n_0, θ, β, A, and ϕ. Most of them are quite flexible, and in our examples, we do not estimate them from real data but specify their values in the same way as in different literature. We will do it in our detailed example.

Finally, let us focus on the estimation of $\pi_{s_1}, \pi_{s_2}, \pi_{s_3}$. At any time t, we have $\pi_{s1} c_t^s c_t^i + \pi_{s2} n_t^s n_t^i + \pi_{s3} = \pi_t$, where π_t is the transmission rate in classic SIR model. Similar to π_r and π_d, the quantity π_t is also available in different data source[2] . To estimate $\pi_{s1}, \pi_{s2}, \pi_{s3}$, we choose two different time spots t_1 and t_2. The first time spot t_1 can be any time between the onset of the spreading of the virus and the first lockdown measure, and the second time spot t_2 must be in a period where a lockdown measure was applied. With the observation of π_{t_1} and π_{t_2}, we have:

$$\pi_{s1} A^2 n_0^2 + \pi_{s2} n_0^2 + \pi_{s3} = \pi_{t_1}, \quad \pi_{s1} A^2 n_0^2 L_{t_2}^2 + \pi_{s2} n_0^2 L_{t_2}^2 + \pi_{s3} = \pi_{t_2},$$

where L_t is an estimation of actual lockdown rate at time t. These two equations are not enough to give us the values of three parameters, we still need one more equation for the purpose. In the case (as happened in the UK) that no different (non-null) lockdown measures have been applied, the third equation is officially not available. So we assume that

$$\pi_{s2} n_0^2 \times \frac{1}{3} \times \frac{1}{6} = \pi_{s3}.$$

[2]HPCC systems covid19: `https://covid19.hpccsystems.com/`

This equation is from the assumption that susceptible people spend about 1/3 of their working hours for other activities related to other types of direct contact, and infectious people spend about half the time of susceptible ones in this type of activity due to the poor health condition. The two proportions 1/3 and 1/6 can be adjusted based on personal experience.

These three equations can give us a good estimation of π_{s1}, π_{s2} and π_{s3}.

1.3.1 Parameter estimation: an example

We take the COVID-19 in the UK as our example, which started in the year 2019. The only lockdown took place on 23 March 2020 and lifted up in July 2020.

For the estimation of π_{s1}, π_{s2} and π_{s3}, we need to specify some other parameters.

According to the starting of the epidemic and lockdown, we take t_1 to be a time in Jan 2020 and t_2 to be some time in April 2020.

The government released Experimental results of the pilot Office for National Statistics (ONS) online time-use study (collected 28 March to 26 April 2020 across Great Britain) [3] compared with the 2014 to 2015 UK time-use study, which reported the working-not-from-home time. According to the study, the average daily time (in minutes) of working not from home is 97.6 in March/April 2020 and 150.0 in 2014/2015, thus we estimate $L_{t2} = 97.6/150 \approx 0.65$.

Also according to ONS, the average actual weekly hours of work for full-time workers from Dec 2019 to Feb 2020 was 36.9 [4] , thereby we set $n_0 = 36.9$. According to the equation $n_0^2\theta = 1$, we set $\theta = 0.00073$.

We follows the setting of some parameters in literature. The mortality rate is set to be 0.6% from [15]. As in [15] , we assume that each infected case takes 18 days on average to either recover or die. Since our model is weekly, we have $\pi_d = 0.006 \times 7/18, \pi_r = 7/18 - \pi_d$. The reproduction number R_0 at time t_1 in Jan 2020 is around 1.95 without control measures [5], and between 0.7 to 1.0 in April 2020 after the lockdown [6][7], we use the middle point 0.85 of this range of R_0 for the calcuation of π_{t_2}. Since in classic SIR model, $R_0 = \beta/\gamma$ where β and γ the infected and recovery transmission rate

$$\pi_{t_1} = 1.95 \times 7/18, \pi_{t_2} = 0.85 \times 7/18.$$

[3]ONS Dataset `https://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/datasets/coronavirusandhowpeoplespenttheirtimeunderlockdown`

[4]ONS, Average actual weekly hours of work for full-time workers:`https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/timeseries/ybuy/lms`

[5]Coronavirus wikipedia: `https://en.wikipedia.org/wiki/Coronavirus_disease_2019`

[6]BBC report on R number: `https://www.bbc.co.uk/news/health-52677194`

[7]The R number in the UK:`https://www.gov.uk/guidance/the-r-number-in-the-uk`

Given a published average annual income [8] 30350 for 52 weeks, we set $A = 15.8172$.

With all quantities involved in the three equations for $(\pi_{s1}, \pi_{s2}, \pi_{s3})$, we get solution

$$\pi_{s1} = 1.244887 \times 10^{-6}, \pi_{s2} = 1.0336 \times 10^{-4}, \pi_{s3} = 0.01759.$$

By the value $n_0 = 36.9$, we take $\theta = 1/(36.9)^2$.

Finally, we copy the value $\phi = 0.8$ from [15].

1.4 Numerical Results

In this section, we present the result of our numerical experiments under the parameter setting in section 3.2. We do experiments to analyze the impact of the optimal lockdown control policy, the policy when different levels of the cost of death are taken into consideration, early exit and late start of the lockdown policy, and finally the smart containment policy. For every experiment, the initial state is $(S, I, R) = (0.9998, 0.0002, 0)$ and the time horizon is 100 weeks.

1.4.1 Optimal Lockdown Control

As Figure 1 (a), (d) (page 20) shows, if there is no lockdown control, i.e. the lockdown rate is constant 1 for all time, then under our parameter setting, around 15% of the population will be infected, 0.3% of the population will die and the peak of infection will be above 0.6% at week 50. Under the optimal lockdown control, the proportion of Infectious people decrease to 5.22×10^{-5} at week 50, then raises to 2.5×10^{-4} at week 100. 0.37% of the population will become infected and 0.0068% of the population will die by week 100. The optimal lockdown policy reduces the peak of infection by 95.8% and reduces the number of deaths by 97.7%. The significant life-saving is associated with a recession. Figure 1 (e) (page 20) shows the aggregate consumption under optimal lockdown policy decreases 20% compares to the no control case at the beginning, but then constantly increases. The average aggregated consumption fall by 6.6% with the optimal lockdown measure. In Figure 1 (f) (page 20), the optimal lockdown rate starts from around 80%, then gradually release to above 95%, the speed of the increase of the lockdown rate first decreases until around week 50, then increase until week 100.

The increase of the infected proportion is because our model has a finite time horizon, and does not take the consequences after a time horizon of 100 weeks into

[8] statista: average full time annual earnings in the uk: `https://www.statista.com/statistics/1002964/average-full-time-annual-earnings-in-the-uk/`

consideration. In the beginning, the aggregated consumption under optimal lockdown control is 20% less but becomes 8.2% more than that of no control in the end. The reason that the optimal lockdown control policy did not cause a severe recession might be that in the no control case, susceptible people will cut back their working hours, as well as their consumption as the infected population increases, and in the optimal lockdown control restricted the infected population so that susceptible people won't cut back their consumption as much. We proved in section 2 that the recovered and Infectious people will work as much time as possible in order to maximize their own utility, but the behaviour of susceptible people is not certain. In the parameter setting of our experiments, the susceptible people almost work as much as possible just as the infected and recovered people do, but slightly reduce their working hours from the upper bound of lockdown constrain near the end of the time horizon, this behaviour may due to the increase of infected proportion, which raises the risk of getting infected for susceptible people.

In general, the optimal lockdown policy saves lives and is more robust in economic recovery, it brings long-term health benefits and economic growth with the cost of a short-term recession.

1.4.2 Cost of Death

In this subsection, we study how the severity of death regarded by the planners affects the optimal lockdown policy. We set the penalty coefficient of death λ in (17) as $0, 10, 20, 50$, which means the death of 1 people is regarded as the loss of $0, 10, 20, 50$ recovered people by the planner. When $\lambda = 0$, it is the same as the original optimal control model.

Our results in Figure 2 (page 21) show that adding a penalty on deaths makes a huge difference, it significantly slows down the increase of the lockdown rate (Figure 2 (f) (page 21)), thus reduces the proportion of deaths in a great extent: $76.7\%, 83.0\%, 87.2\%$ respectively (Figure 2 (d) (page 21)), and avoid the substantial rise of the infectious population (Figure 2 (a) (page 21)), these are beneficial in terms of the mental impact in the society as low deaths and infection amount release the pressure on both people in the society and the planner. As the penalty coefficient increases, the optimal policy becomes constantly more strict. The relation of the death penalty coefficient and the result optimal control rate is below linear. As Figure 2 (f) (page 21) shows, despite that optimal lockdown policy with different death penalty coefficient starts with quite different lockdown rates:$0.62, 0.56, 0.46$ for penalty coefficient $10, 20, 50$ respectively, they quickly become close. At the end of control, the aggregate consumption, as well

as the lockdown rate of optimal policy with the death penalty is extremely close to the one without the death penalty. Compare to the original optimal lockdown policy, there is a slight recession when adding penalty on number of deaths: the average aggregate consumption decreases by $3.3\%, 4.5\%, 5.5\%$ for penalty coefficient $10, 20, 50$ respectively (Figure 2 (e) (page 21)).

Although the lockdown control policy with or without the death penalty becomes close from the middle to the end of the control, in the latter case, the infectious population does not rise as it in the former case. This is because that the lockdown policy with the death penalty suppresses the infectious population to a much lower level than the lockdown policy without the death penalty, thus the infectious population grows slower as the lockdown rate increases.

In general, considering the cost of deaths leads to a more conservative lockdown control policy, it saves much more lives at the cost of a short-term recession.

1.4.3 Cost of Early Ending of Lockdown Control Policy

Practically, policymakers may under the intense pressure of economic loss that forces them to end the containment policy in the middle of the pandemic. In this subsection, we discuss the consequences of doing so. As we see in section 4.1, the infected population reaches the bottom at the week 50, which may seem to be a good time spot to end the lockdown policy.

Our results in Figure 3 (e) (page 22) shows that there is an instant bounce of consumption right after the end of lockdown control, but this would cause the instant rise of infectious population (Figure 3 (a) (page 22)), and at the end, the infectious population 72 times larger than that of week 50. The burst of infection would result in a recession of 10.8% from the peak at the end (Figure 3 (e) (page 22)).

So, ending the lockdown policy prematurely may not bring long-term economic benefit and what's worse is, it would result in a substantial additional number of deaths. Therefore we suggest that policymakers avoid terminating the lockdown policy during the pandemic in pursuit of only a short-term economic benefit.

1.4.4 Cost of Start the Lockdown Control Policy Late

Policymakers could also face the situation that there are things that prevent them from taking the lockdown measure in the early stage of the pandemic.

Our results Figure 4 (f) (page 23) show the optimal lockdown policy that starts at week 13 (around 3 months later). Compare to the optimal lockdown policy that starts

at week 0 that starts with the lockdown rate 0.8, the late started optimal lockdown policy starts with a stricter constrain rate of 0.73. Although the lockdown rate of the late started lockdown policy constantly increases, it is always less than the original optimal lockdown policy. The late start causes a slight stronger recession (Figure 4 (e) (page 23)): the average aggregated consumption reduces 1% and a substantial rise of deaths (Figure 4 (d) (page 23)): the number of deaths rises 84.8% by week 100.

In general, It is the earlier the better to start the lockdown control policy, and despite that the late start of lockdown policy brings additional loss, it is much better than applying no containment policy or abandon it too early.

1.4.5 Vaccination

Vaccination is an effective method of preventing infectious diseases. We now involve vaccination in SIR model. Assume that at each time period, fix amount of susceptible people: δ_v of the starting population get vaccination that could prevent them from getting COVID-19 and assume governments to afford the cost of vaccination for people. Once susceptible people get vaccination, they are regarded as recovered. Thus the objective value of susceptible people become:

$$\begin{aligned} J^s(c^s_\cdot, n^s_\cdot; t, X_t, L_\cdot) &= u(c^s_t, n^s_t) + \beta(1 - \frac{\delta_v}{S_t})\tau_t J^{i*}(t+1, L_\cdot) \\ &\quad + \beta(1 - \frac{\delta_v}{S_t})(1-\tau_t) J^s(c^s_\cdot, n^s_\cdot; t+1, X_{t+1}, L_\cdot) \\ &\quad + \beta \frac{\delta_v}{S_t} J^{r*}(t+1, L_\cdot) \end{aligned} \tag{1.24}$$

With the cost of vaccination, denote as p, the optimal control problem of policymakers become:

$$\max_{L_\cdot} \quad J^0(L_\cdot; t, X_t) = \sum_{t=t_0}^{T} \beta^{t-t_0} \left[S_t u(c^{s*}_t, n^{s*}_t) + I_t u(c^{i*}_t, n^{i*}_t) + R_t u(c^{r*}_t, n^{r*}_t) - p\delta_v \right] \tag{1.25}$$

We set $\delta_v = 1/104$ in simulation. Results on Figure 7(a),(b),(d)(page 26) shows that vaccination could eliminate the epidemic without a rebound of infection and reduce the number of deaths compare that without vaccination. Figure(e),(f) shows that vaccination reduces the severity of recession and leads to a less strict optimal lockdown control.

1.4.6 Smart Lockdown Control Policy

In the lockdown control policies we studied so far, the government chooses the same lockdown rate for all three kinds of people (susceptible, infectious, and recovered). In this subsection, we consider the smart containment, by which means the policymaker directly chooses working hours for all three kinds of people with the same objective function as previous models. There is no need to apply any lockdown on recovered people because their utility reaches the maximum as their working hour is at the maximum and they do not affect the utility or the transition of susceptible and Infectious people. Our results show that in the smart lockdown control policy, Infectious people almost do not work at the beginning, but then the planner gradually increases their working hours as the infected population decreases rapidly, and susceptible people work almost without fear of becoming infected. Figure 5 (page 24) shows that compare to the previous optimal lockdown control policy, the smart lockdown policy is much better, since it reduces the number of deaths to a great extent, and almost avoids the recession because the proportion of Infectious people is extremely small. The implement of a smart lockdown control policy requires the planners to know the status of all people and have control over their working hours. In reality, the knowledge of people's status needs measures such as medical testing and rely on the accuracy of testing. Our results suggest that these measures and information that are helpful for taking smart lockdown policy are beneficial for social welfare.

1.4.7 View of Reproduction Number

The reproduction number (R_0) is now a basis for some governments to make decisions in reaction to the pandemic. We present the R_0 of lockdown polices in all our experiments in Figure 6 (page 25). The R_0 of smart lockdown policy is much smaller than that of all other lockdown policies. The R_0 of lockdown policies with the same lockdown rate for all three kinds of people behave similarly to their lockdown rate whereas in the no control case, its R_0 decreases constantly and the R_0 of smart lockdown policy behaves similar to the lockdown rate of Infectious people. This is because the behaviour of all three kinds of people is in accordance with the lockdown rate in optimal lockdown control policies, while the behaviour of susceptible people varies if there is no control, and in the smart lockdown case, the susceptible and recovered people almost remain the lockdown rate as constant 1 in the whole control process, thus its R_0 behaves in accordance with the lockdown rate of Infectious people. Notice

that although the R_0 of in the no control case decreases below 1, and the R_0 of lockdown control policies with or without the death penalty increase over 1, policies with lockdown control are much better than that without control as analyzed in previous subsections. We, therefore, suggest that whether R_0 is larger or less than 1 can not be the only foundation for planners to make judgements or decisions on the current situation.

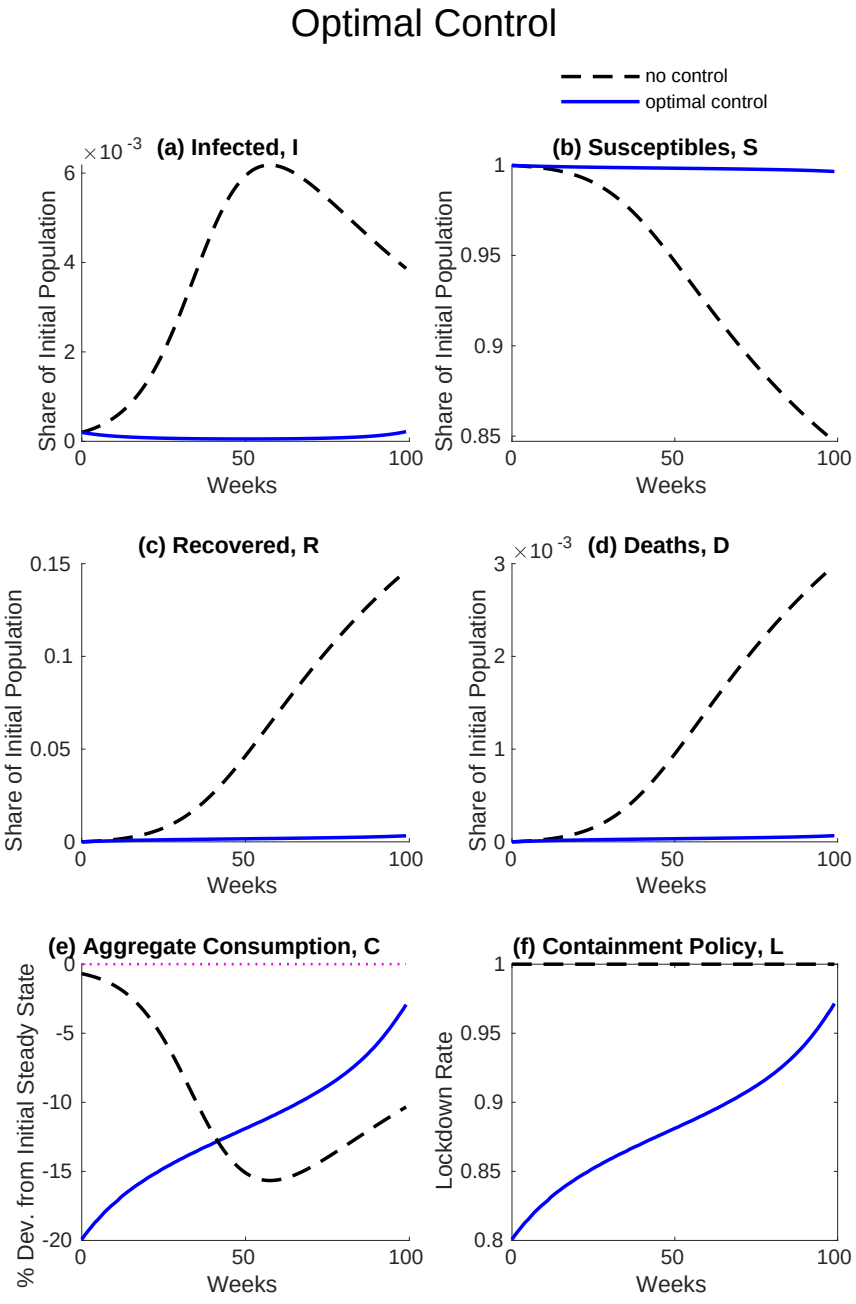

Figure 1.1: Optimal Control

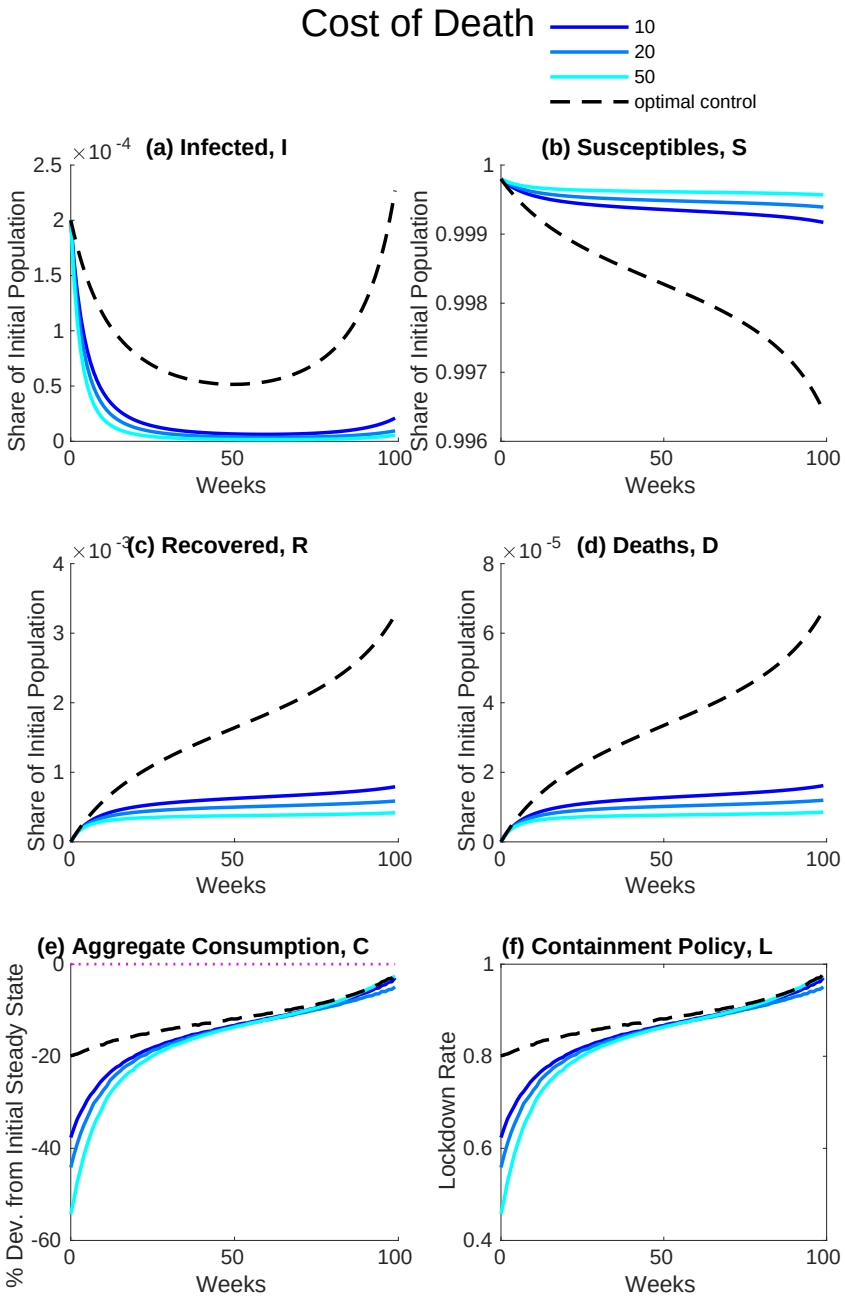

Figure 1.2: Cost of Death

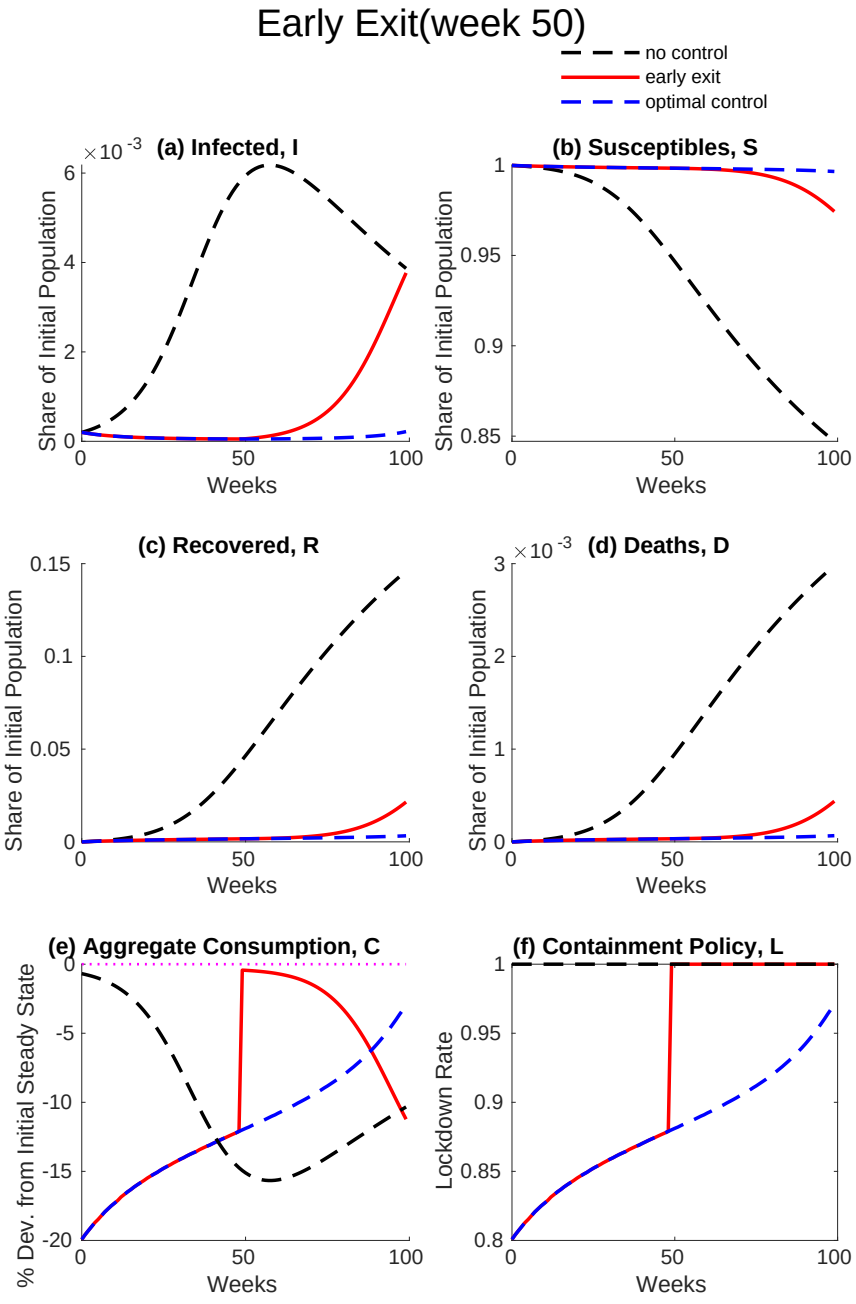

Figure 1.3: Early Exit

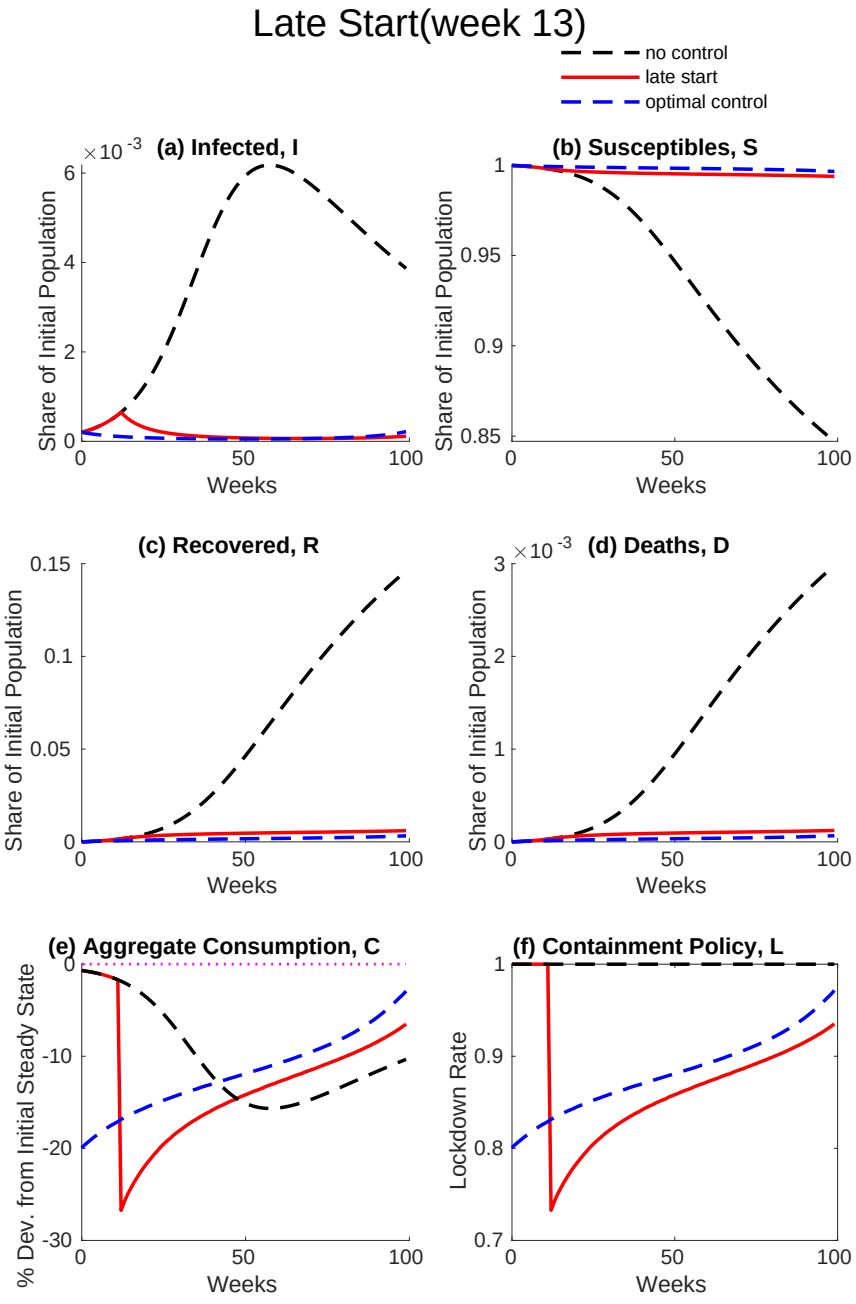

Figure 1.4: Late Start

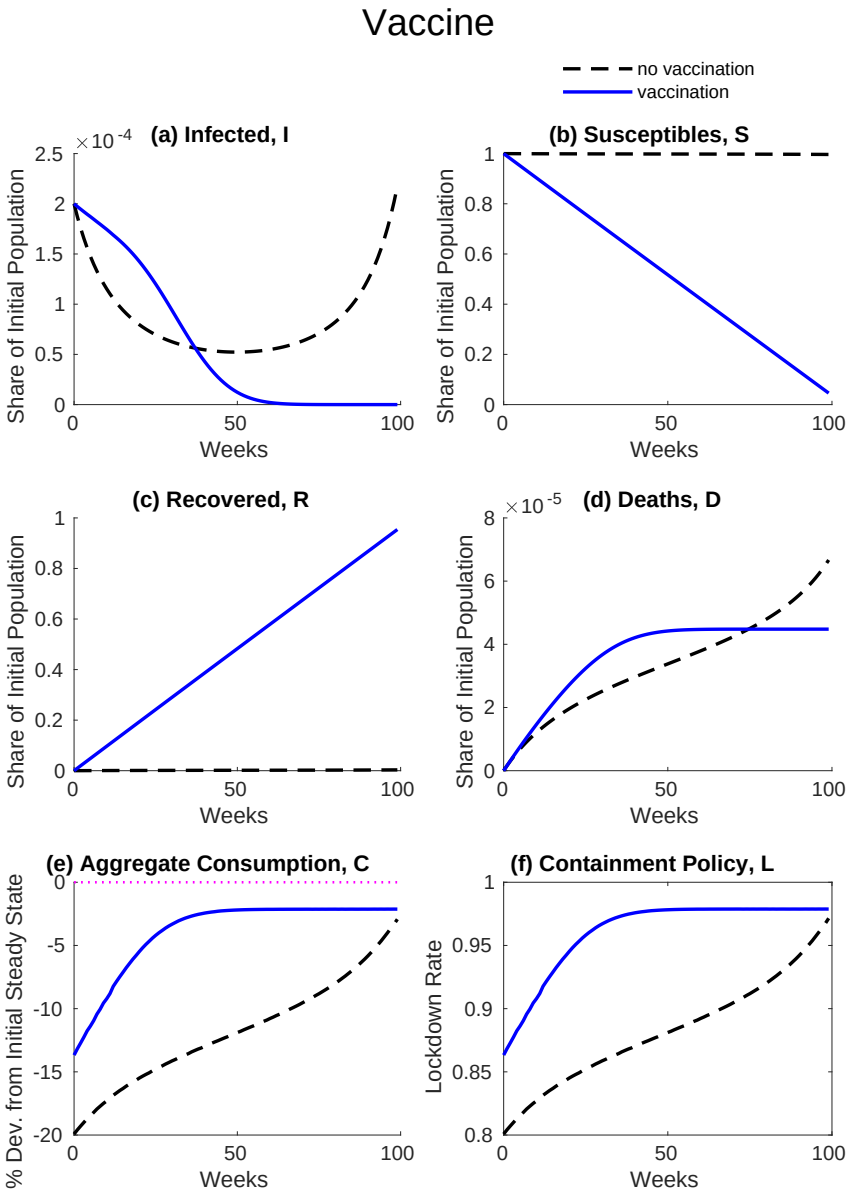

Figure 1.5: Vaccine

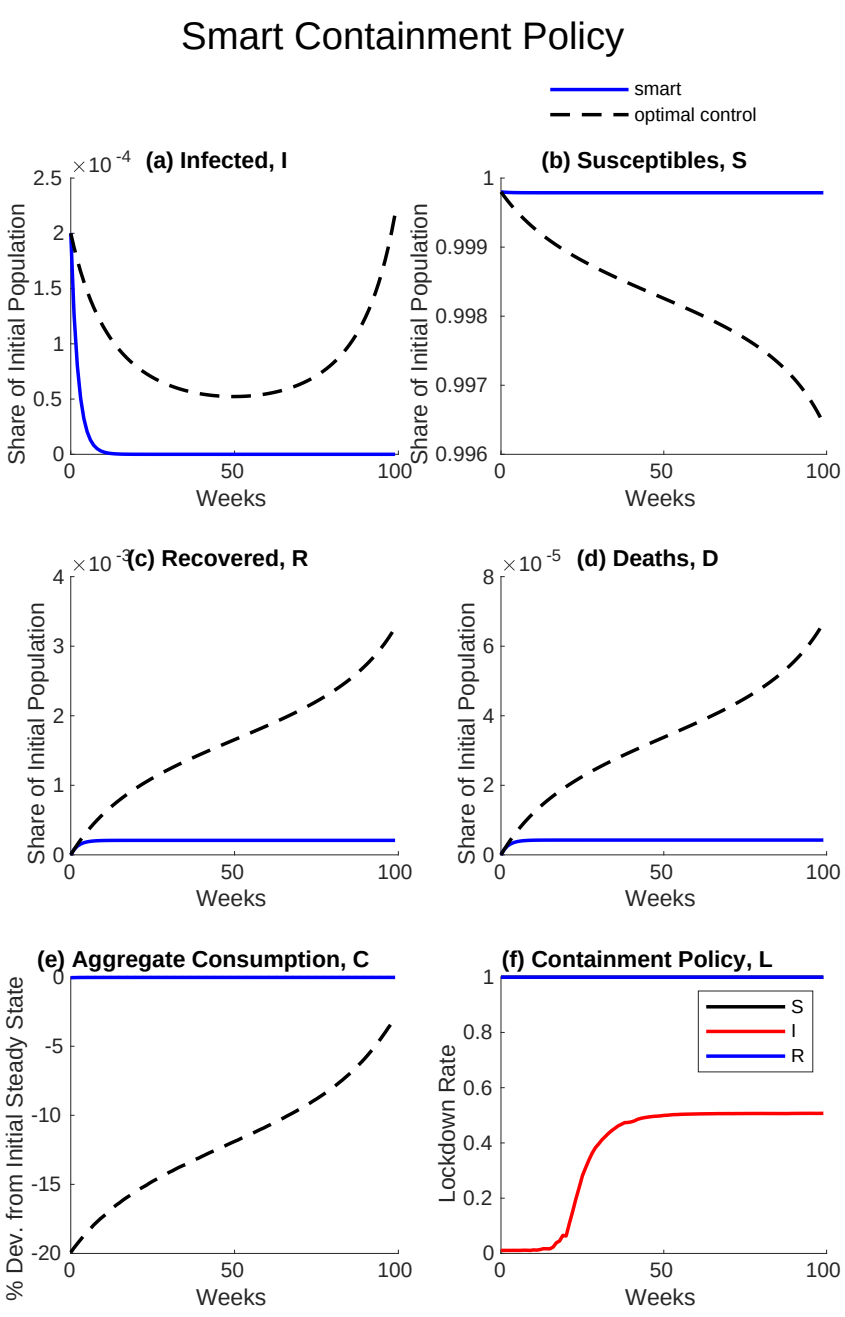

Figure 1.6: Smart Control

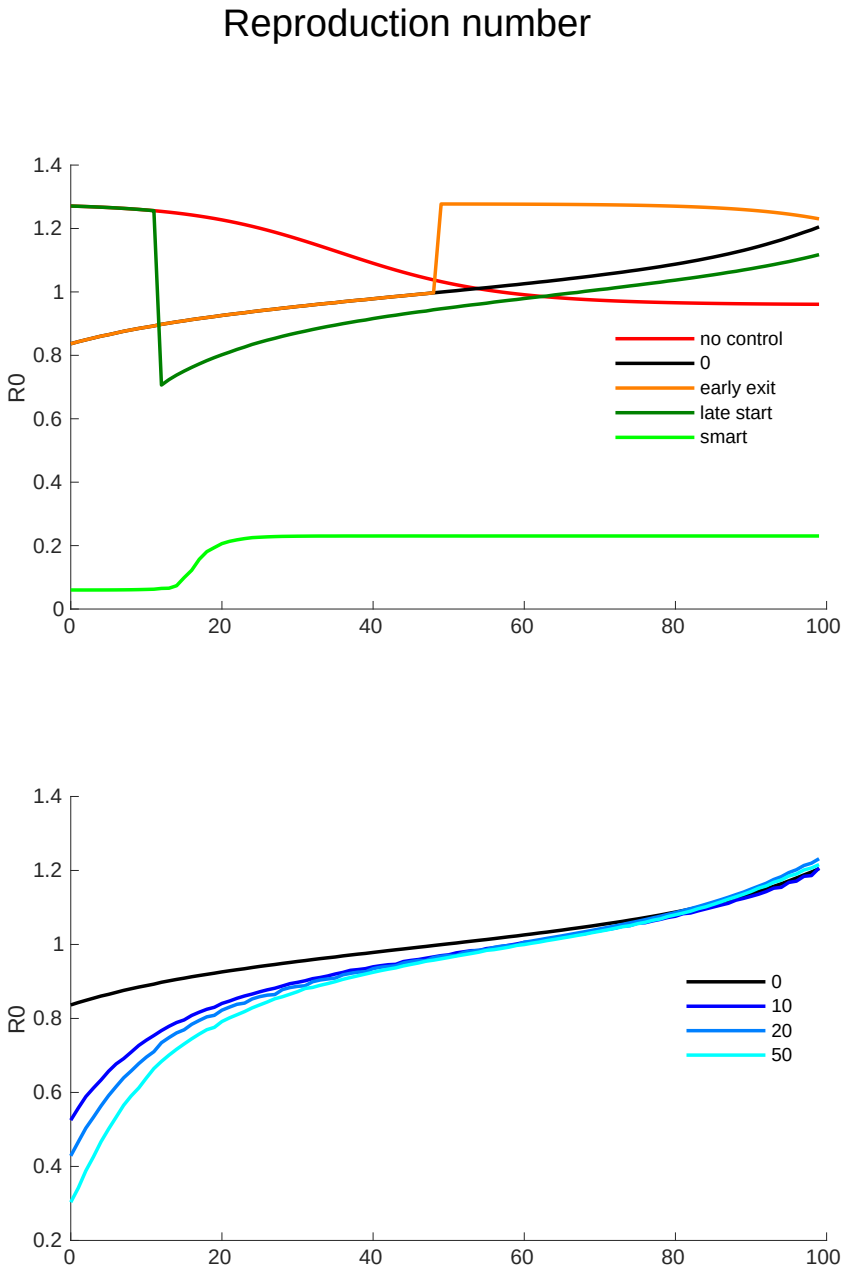
Reproduction number
R0
no control
0
early exit
late start
smart
R0
0
10
20
50

Figure 1.7: R0

Chapter 2

Imitation Learning in Finance

2.1 Introduction

Imitation learning, also referred as learning from demonstration, is a prominent branch of machine learning that focuses on reproducing expert behaviors by leveraging expert demonstrations in the form of state-action trajectories. Similar to reinforcement learning, imitation learning solves optimal control problems in a data-driven way, but differently from reinforcement learning which rely on optimizing rewards through trial and error, imitation learning directly learns from observed examples provided by experts,the emphasis is on imitating expert actions rather than designing reward functions. In recent years, imitation learning methods have shown success in robotics[7] automated driving[12], etc. The success of imitation learning lies in its ability to leverage expert demonstrations to learn intricate behaviors and decision-making strategies. By learning from experienced individuals, the agent can bypass the need for exhaustive trial and error and directly acquire high-quality policies. This makes imitation learning an attractive approach in situations where expert knowledge is available, or when manually designing optimal policies is difficult or time-consuming.

The applications of imitation learning methods in financial markets is a relatively new and emerging research area. Traditional rule-based strategies or models may have limitations in capturing the complexities and dynamics of financial markets. By leveraging imitation learning, we can harness the expertise and decision-making processes of investors to develop systems capable of making informed investment decisions.

2.2 Background

For a Markov decision process(MDP), $(\mathcal{S}, \mathcal{A}, R, P, p_0, \gamma, T)$ where S is the state space, A is the action space, $\mathcal{R} : (\mathcal{S} \times \mathcal{A}) \to \mathcal{R}$ is reward, without losing generality, we consider the reward $0 \leq r(s,a) \leq 1 \forall (s,a) \in (\mathcal{S} \times \mathcal{A})$,$P(s,s',a)$ is transition matrix, $p_0(s)$ is the initial state distribution,discounted factor $0 < \gamma < 1$, for infinite horizon, and $\gamma = 1$ for finite horizon T. Occupancy measure of policy, i.e the distribution of state-action pair (s,a) by following policy π, $\rho_\pi(s,a) = \frac{1}{T}\sum_{t=1}^{T} p_t(s,a)$ for finite horizon and $\rho_\pi(s,a) = \sum_{t=1}^{\infty} \gamma^t p_t(s,a)$ for infinite horizon. A stationary policy $\pi(a|s)$ denotes the action distribution given state s, $\pi_E(a|s)$ is the expert policy. Value function $V_\pi = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, a_t\right) \mid a_t \sim \pi\left(\cdot \mid s_t\right)\right]$ for infinite horizon and $V_\pi = \mathbb{E}\left[\sum_{t=0}^{T} r\left(s_t, a_t\right) \mid a_t \sim \pi\left(\cdot \mid s_t\right)\right]$ for finite horizon.

In imitation learning, we have collected expert demonstrations $\{\tau_i := (s_t, a_t)_{t=1}^{T}\}_{i=1}^{N}$, where N is the number of expert demonstrations. With the collected expert demonstrations, the goal of imitation learning is to learn an optimal policy π^* s.t.

$$\pi^* = \arg\min D(q(\boldsymbol{\phi}), p(\boldsymbol{\phi})) \tag{2.1}$$

where $p(\phi)$ and $q(\phi)$ are two distributions of features ϕ, features can be states and actions of the MDP or any other measurements,$D(p,q)$ is a similarity measure between p, q[49]

2.3 Literature

There are mainly two kinds of approaches for imitation learning problems: 1) Inverse reinforcement learning that learns a specific reward function, then learn the expert policy by reinforcement learning methods. 2) Learn policy directly. In this section, we summarise the representative and commonly used methods in these two categories, as well as other variations of imitation learning problem setting. Finally we focus on applications of imitation learning methods in financial markets.

2.3.1 Inverse Reinforcement Learning

In inverse reinforcement learning, commonly assumes the expert follows a Markov decision process(MDP) for decision making, the learner tries to recover the reward function R from demonstrations generated by expert policy. Then reinforcement learning methods could be applied to learn the expert policy based on the learned reward function. One big challenge of applying reinforcement learning to real-world

problems is the design of reward function(For example, if we are going to teach a robot to pick up an apple from the table, it would be difficult to hand-craft the reward function in multi-dimensional space of the robot's states), inverse reinforcement learning could server as a solution to this problem. Reward function could represent agents' preference. In financial applications, knowing the utility of a counterparty may be useful in bilateral trading, e.g. over-the-counter (OTC) trades in derivatives or credit default swaps. Moreover, the learned reward function could be transferable to other similar decision-making problems.

Inverse reinforcement learning methods has the problem of reward ambiguity[10], Ng et.al[45] proved that: $\forall\phi : S \rightarrow R$ the optimal policy remains optimal under transformation:

$$\hat{r}\left(s, a, s'\right) = r\left(s, a, s'\right) + \gamma\phi\left(s'\right) - \phi(s) \tag{2.2}$$

Although given demonstrations of actions for the same reward under two distinct discount factors, or under sufficiently different environments, the unobserved reward can be recovered up to a constant[40], these conditions are often not satisfied in real-world problems. Therefore external constrains need to be added to obtain unique reward. Feature matching[46] assume reward function is linear expansion of some features and maximize the margin between the optimal policy and others. The maximum entropy principle is often applied for IRL in recent studies. Maximum Entropy Inverse Reinforcement Learning(MaxEnt[74] learns the policy that maximize the entropy (... that has the maximum entropy), this leads to the policy that follows Boltzman distribution:

$$\pi_\theta(a|s) = \frac{1}{Z(s)} e^{r_\theta(s,a)} \tag{2.3}$$

where $Z(s) = \int_a e^{r(s,a)}$ is the partition function. When the environment dynamics is known, $Z(s)$ can be computed by dynamic programming[74]. One drawback of MaxEnt is that it inference problems with i.i.d data, but the policy at time t should not depend on future trajectory. Maximize Causal Entropy Inverse Reinforcement Learning[73] fixes this by learning the reward s.t. the optimal policy has the maximum the causal entropy. For environment with unknown dynamics, importance sampling in cases where states dimension are high[20]. Model-based methods that learns the transition model could also be applied[49]. These kind of method is data-efficient in terms of expert trajectory, but model learning could be difficult.

2.3.2 Learning Policy

Unlike inverse reinforcement learning that learns the reward function before learning the expert policy, we could learn the mapping from states to actions directly. It is natural form imitation learning as a supervised learning problem in which the policy is obtained by solving a single regression problem. This kind of method is referred as Behaviour Cloning(BC)[5]. BC minimizes the discrepancy of action distribution between the expert and the learner policy. The discrepancy measure quadratic loss[36], neural networks[61], Kullback-Leibler(KL) divergence etc. Then the imitation learning problem can be formulated as

$$\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_{\pi_{\mathrm{E}}}} \left[D_{\mathrm{KL}} \left(\pi_{\mathrm{E}}(\cdot \mid s), \pi(\cdot \mid s) \right) \right] \tag{2.4}$$

Data Aggregation(DAgger)[57] is the extension of BC in active imitation learning setting in which the learner can collect expert demonstrations under the state distribution generated by the policy of the learner. Although BC methods may be able to obtain policies that match expert demonstrations well, they have two concerns. 1) Covariate shift[Ross 2010]. Supervised imitation learning can be challenging when demonstrations do not cover the states that the learner encounters. 2) Compounding errors[Ross 2011]. Small one-step error would accumulate with time, which will lead to significant different state-action distribution between the expert and the learner policy.

One recent trended method that could alleviate the above problems of BC is the Generative Adversarial Imitation Learning(GAIL)[34]. Like BC, GAIL focus on the problem of learning the optimal policy without learning the reward. GAIL bypasses the direct reinforcement learning problem by replacing the specific reward function with a reward function defined on a space of all admissible reward functions. The imitation learning problem is:

$$\begin{aligned}
\mathrm{RL} \circ \mathrm{IRL} \left(\pi_E \right) &= \min_{\pi} -H(\pi) + \max_{r(s,a) \in R} -\psi(r) \\
&\quad + \mathbb{E}_{\rho_\pi(s,a)}[r(s,a)] - \mathbb{E}_{\rho_E(s,a)}[r(s,a)] \\
&= \min_{\pi} -H^{\mathrm{causal}} \left(\pi \right) \\
&\quad + \max_{r(s,a) \in R} \left(\sum_{s,a} \left(\rho_\pi(s,a) - \rho_E(s,a) \right) r(s,a) - \psi(c) \right) \\
&= \min_{\pi} -H(\pi) + \psi^\star \left(\rho_\pi - \rho_E \right)
\end{aligned} \tag{2.5}$$

where $\psi(r)$ is the convex regularization function, $\psi^*(r)$ is its Fenchel conjugate and $H(\pi) \triangleq \mathbb{E}_\pi[-\log \pi(a \mid s)]$ is the causal entropy of π. The original GAIL chooses $\psi(r)$ to be the Jensen-Shannon(JS)divergence[23].

The policy learned by GAIL is an analogy to the generator in Generative Adversarial Networks(GANs[29]), the discriminator $D(s, a)$ distinguishes the trajectories generated by the learner's policy π and the expert policy π_E. As shown in Dixon et.al[14], 2.5 can be written as:

$$\min_{\pi \in \Pi} \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} \mathbb{E}_{(s,a) \sim \rho_\pi}[\log(D(s, a))] + \mathbb{E}_{(s,a) \sim \rho_{\pi_{\mathrm{E}}}}[\log(1 - D(s, a))] - H(\pi) \qquad (2.6)$$

The adversarial training process of GAIL is similar to that of GANs, which is to update policy network(the generator) and the discriminator network parameters iteratively. For each iteration of the policy network, it takes one policy gradient step with cost function $\log D(s, a)$ in reinforcement learning. Note that this is not the cost function of the original imitation learning problem:$D(s, a) \rightarrow 1/2 \forall (s, a) \in \mathcal{S} \times \mathcal{A}$, and GAIL does not learn the reward function of the expert.

Trust Region Policy Optimization(TRPO)[59] or Proximal Policy Optimization(PPO)[60] is applied to avoid drastic policy change due to the noise of policy gradient estimation.

GAIL is computationally more efficient than IRL methods. However, the training of GAIL and other generative adversarial methods can be unstable and has convergence problems.

GAIL has various variations and generalizations. In cases that the expert demonstration has latent factors(e.g the expert demonstration contains data collected from several experts), Information Maximizing Generative Adversarial Imitation Learning(i.e InfoGAIL)[43] infers the latent structure of expert demonstrations by adding the mutual information between latent codes and trajectories as a regulator to the original GAIL objective, by which the learner could imitate the expert behavior while separating different kinds of expert behaviors. The mutual information regulator could be approximated by introducing a variational lower bound. In the original GAIL and most of it's variations learn the policy that maximize the expected total reward of trajectories, but not consider higher order moments of the reward. For applications that risk is crucial, for example, financial applications such as portfolio trading, risk-sensitive generative adversarial imitation learning(RS-GAIL)[42] adds a constrain to the original GAIL objective that learner has higher risk than the expert under risk measures such as conditional value at risk(CVaR).

f-VIM(f-Divergence Variational Imitation)[38] gives a framework that formulate the imitation learning problem as the problem of minimizing the f-divergence between

the expert and learner state-action distribution. GAIL is a special case when use JS divergence in f-VIM besides the entropy regulator. Since JS and KL divergence have a mode-covering behaviour that interpolate across modes, Ke et al.[38] chooses the reverse KL divergence which has a mode-seeking behaviour. Instead of applying a specific f-divergence to measure the discrepancy between the expert and the learner behaviours as GAIL and its above variations, f-GAIL[72] learns the best f-divergence under which the discrepancy between the expert and the learner behaviours is the smallest among all feasible f-divergence.

As mentioned before, GAIL focus on recovering the optimal policy without learning the reward. AIRL[22] is an extension of of GAIL that is able to learn both the expert policy and it's reward. AIRL sets a specified form of discriminator function:

$$D_\theta(s,a) = \frac{\exp\left(f_\theta(s,a)\right)}{\exp\left(f_\theta(s,a)\right) + \pi(a \mid s)} \tag{2.7}$$

where $\pi(a|s)$ is the learner policy and $f_\theta(s,a)$ is the function to learn. At the optimum, $f_\theta(s,a)$ equals the advantage function which can be utilized to recover one-step disentangled rewards. Ghasemipour et al.[25] proved that AIRL is equivalent to minimize the reverse KL divergence between the state-action distribution of the expert and the learner. f-DIVERGENCE MAX-ENT IRL(f-MAX)[26] generalizes AIRL for a general f-divergence.

2.3.3 Other Setting

Imitation learning methods we discussed above assume the environment is modeled as a MDP, and the expert demonstration data contains complete information about states and actions. In real-world problems, complete states and actions are unavailable or expensive.

In problems with partial observable states(e.g imitate a trader's trading behaviour with stock price data, but the trader utilizes information other than the stock price for trading), the decision-making process of the expert is modeled as Partial Observable Markov Decision Process(POMDP). POMDP algorithms use the concept of belief states, i.e. the probability distribution over the current states. Chio and Kim[11] extends feature expectation IRL[46] to POMDP setting. Bogert et al.[8]extends MaxEnt by applying Expectation Maximization(EM). Belief-module Imitation Learning (BMIL)[24] learns the belief representation of states using historical states observations then combine with GAIL.

Imitation Learning Form Observation(IfO) learns without complete actions. For example, robots learning actions from video demonstrations rather than direct state-action pairs. IfO algorithms recover actions from demonstrations, then apply standard imitation learning methods or approaches in reinforcement learning. The state-action transaction model learning can mainly be classified as inverse dynamics models that learns the mapping from $(S_t, S_{t+1})$to A_t[31]and forward dynamics models that learns the mapping from (S_t, A_t) to S_{t+1}[50].

2.3.4 Applications in Finance

In this subsection, we summarize current applications of imitation learning in finance. Although the literature contains relatively few published studies on this topic so far, the increasing availability of large financial datasets presents new opportunities to apply data-driven imitation learning techniques. With the growth of electronic trading platforms producing high-frequency transaction data and order flow information, there is strong potential for imitation learning methods to gain traction for modeling and inference tasks in financial markets. We will consider policy imitation and reward inference of two kinds of problems: single market agent and collective behaviour of all agents.

Yang et al.[69] addressed the problem of trading strategy identification given historical Limit Order Book(LOB) trading records. They compared Gaussian Process Inverse Reinforcement Learning(GPIRL)[53] in which the reward is modeled as a Gaussian Process with Linear IRL(ILRL)[46] which assumes the expert policy is deterministic and found rewards learned by GPIRL as features result in better performance in trading strategy classification. The learned rewards have the potential to be informative features of perceived goals of traders.

Investor sentiment is commonly viewed to have strong impact on prices of securities. Yang et al.[68] explores the link between a proxy to investor sentiment and future market movements by applying GPIRL. They regard the the investor sentiment(Thompson Reuters' News Sentiment) as the collective action of market participants and market stock returns and volatilities as states in MDP. The objective of this kind of modeling is to find useful high-level features that can be used to construct better predictive models for equity returns. Experiments have shown that an adaptive trading system based on learned reward functions performed better than the original investor sentiment proxy.

In marketing literature, the inverse optimal control methods to learn the customer utility is referred as structural models[44]. These models assume customers

are rational agents that maximize their utilities and typically specify a model for the consumer utility, and then estimate such a model using methods of dynamic programming. Halperin[14] applied MaxEnt-IRL[74] with specially chosen reference action distribution to infer customer preferences, this method is computationally more efficient than structural model approaches.

Roa-Vicens et al.[56] applied AIRL and GAIL to recover trading agents polices trained by reinforcement learning algorithms in an latent-space modeled LOB environment using stock returns as rewards. Their experimental results show that AIRL outperforms GAIL in total returns, which may indicate AIRL is more robust than GAIL. Notice that there is large gap between the total return of experts and learners, this may due to the training and testing period contain significant variablility of price and volatility levels, which is common in securities price time series.

Roa-Vicens et al.[55] made an attempt to learn the LOB dynamic in a simplified LOB environment modeled by a MDP using three different IRL algorithms: MaxEnt, Bayesian Neural Network(BNN) IRL and GPIRL. While all three algorithms are able to recover the linear expert reward, BNN IRL and GPIRL outperforms MaxEnt for recovering non-linear expert rewards.

Robo-advisors are a class of financial advisers that provide online financial advice or investment management with human intervention[65]. Robo-advising algorithms learn the client's risk preference through interactions with the client or by viewing the client's historical investments. Yu and Dong[70] utilized inverse optimization techniques under the multi-period mean-variance framework to infer time-varying risk preference of traders. Imitation learning methods could be a model free alternative to this problem.

2.4 Analysis of BC and GAIL

In this section, we present some analysis of error bounds for behavioural cloning, GAIL and its variations, as well as the generalization ability of GAIL. We consider time-invariant policies with the finite horizon setting.

2.4.1 Error Bounds for BC and GAIL

For behavioural cloning, Syed et al. and Ross et al.[63][57] established error bounds in the case of deterministic expert polices. Different from the above works, we proved the following theorem that considers stochastic expert polices.

Theorem 4. *Given expert policy π_E for MDP with finite horizon H and an imitated policy with an imitated policy π with behavioural cloning objective $\mathbb{E}_{s\sim d_{\pi_{\mathrm{E}}}}\left[D_{\mathrm{KL}}\left(\pi_{\mathrm{E}}(\cdot \mid s), \pi(\cdot \mid s)\right)\right] \leq \epsilon$, where $d_E = \sum_{t=1}^{H} d(s,a)$, then*

$$V_{\pi_{\mathrm{E}}} - V_{\pi} \leq H^2\sqrt{2\epsilon} \tag{2.8}$$

Proof. $\forall t \in 1, ..., H$, define policy $\tilde{\pi}^t = \{\pi_E^1, ..., \pi_E^t, \pi^{t+1}, ..., \pi^H\}$

then we have,

$$V\left(\pi_E\right) - V(\pi) = \sum_{\tau=1}^{H} V\left(\widetilde{\pi}^{\tau}\right) - V\left(\widetilde{\pi}^{\tau-1}\right) \tag{2.9}$$

$$V\left(\tilde{\pi}^{\tau}\right) - V\left(\tilde{\pi}^{\tau-1}\right) = \sum_{t=\tau}^{H} \mathbb{E}_{\widetilde{\pi}^{\tau}}\left[\mathbf{r}_t\left(s_t, a_t\right)\right] - \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[\mathbf{r}_t\left(s_t, a_t\right)\right] \tag{2.10}$$

$\forall t \in \{1, ..., H\}$ and $\forall \tau \geq t$

$$\begin{aligned}
\mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[\mathbf{r}_t\left(s_t, a_t\right)\right] &= \mathbb{E}_{s_\tau \sim f^{\tau}_{\pi_E}}\left[\mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[\mathbf{r}_t\left(s_t, a_t\right) \mid s_\tau, a_\tau\right]\right] \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} f^{\tau}_{\pi_E}(s)\pi(a \mid s)\mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[\mathbf{r}_t\left(s_t, a_t\right) \mid s_\tau = s, a_\tau = a\right] \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} f^{\tau}_{\pi_E}(s)\pi(a \mid s)\mathbb{E}_{\pi}\left[\mathbf{r}_t\left(s_t, a_t\right) \mid s_\tau = s, a_\tau = a\right]
\end{aligned} \tag{2.11}$$

By similar decomposition,

$$\mathbb{E}_{\widetilde{\pi}^{\tau}}\left[\mathbf{r}_t\left(s_t, a_t\right)\right] = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} f^{\tau}_{\pi_E}(s)\pi_E(a \mid s)\mathbb{E}_{\pi}\left[\mathbf{r}_t\left(s_t, a_t\right) \mid s_\tau = s, a_\tau = a\right] \tag{2.12}$$

subtracting 2.11 from 2.12, we have

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\pi}^{\tau}}[\mathbf{r}_t\left(s_t, a_t\right)] - \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[\mathbf{r}_t\left(s_t, a_t\right)\right] \\
&\quad \leq \sum_{s\in\mathcal{S}} f^{\tau}_{\pi_E}(s) \sum_{a\in\mathcal{A}} \mathbb{E}_{\pi}\left[\mathbf{r}_t\left(s_t, a_t\right) \mid s_\tau = s, a_\tau = a\right]\left(\pi_E(a \mid s) - \pi(a \mid s)\right)
\end{aligned} \tag{2.13}$$

Since $0 \leq r(s,a) \leq 1$,

$$\begin{aligned}
\mathbb{E}_{\widetilde{\pi}^{\tau}}\left[\mathbf{r}_t\left(s_t, a_t\right)\right] - \mathbb{E}_{\widetilde{\pi}^{\tau-1}}\left[\mathbf{r}_t\left(s_t, a_t\right)\right] &\leq \sum_{s\in\mathcal{S}} f^{\tau}_{\pi_E}(s) \sup_{g:\mathcal{A}\to[0,1]} \sum_{a\in\mathcal{A}} g(a)\left(\pi_E(a \mid s) - \pi(a \mid s)\right) \\
&= \sum_{s\in\mathcal{S}} f^{\tau}_{\pi_E}(s)\,\mathrm{TV}\left(\pi_E(a \mid s), \pi(a \mid s)\right) \text{ (TV: Total Variation)} \\
&= \mathbb{E}_{s\sim f^{\tau}_{\pi_E}}\left[\mathrm{TV}\left(\pi_E(a \mid s), \pi(a \mid s)\right)\right]
\end{aligned} \tag{2.14}$$

$$
\begin{aligned}
V\left(\pi_E\right)-V(\pi) &\leq H \sum_{\tau=1}^{H} \mathbb{E}_{s \sim f_{\pi_E}^{\tau}}\left[\mathrm{TV}\left(\pi_E(a \mid s), \pi(a \mid s)\right)\right] \\
&=H^2 \mathbb{E}_{s \sim f_{\pi_E}}\left[\mathrm{TV}\left(\pi_E(a \mid s), \pi(a \mid s)\right)\right]
\end{aligned}
\tag{2.15}
$$

By Pinsker Inequality[58],

$$
\begin{aligned}
V\left(\pi_E\right)-V(\pi) &\leq H^2 \mathbb{E}_{s \sim f_{\pi_E}}\left[\sqrt{2 \,\mathrm{KL}\left(\pi_E(a \mid s), \pi(a \mid s)\right)}\right] \\
&\leq H^2 \sqrt{2 \mathbb{E}_{s \sim f_{\pi_E}}\left[\mathrm{KL}\left(\pi_E(a \mid s), \pi(a \mid s)\right)\right]}=H^2 \sqrt{2 \epsilon}
\end{aligned}
\tag{2.16}
$$

□

Next, we derive the error bound for f-VIM, which is a generalization of the original GAIL, it turns to GAIL when choose $f(x) = \ln(x)$

Theorem 5. *Given an expert policy π_E and an imitated policy π, s.t. $D_f(\rho_\pi, \rho_{\pi_E}) \leq \epsilon$(which can be achieved by f-VIM), f is differentiable up to order 3 at $x = 1$ and $f''(1) > 0$, then we have $V_{\pi_E} - V_\pi \leq H\sqrt{\frac{2}{f''(1)}\epsilon}$*

Proof.

$$
\begin{aligned}
V_{\pi_E}-V_\pi &= H(\mathbb{E}_{\pi_E} r(s,a)-\mathbb{E}_\pi r(s,a)) \\
&\leq H \sum\left|\rho_{\pi_E}(s,a)-\rho_\pi(s,a)\right| \\
&= H D_{TV}(\rho_\pi, \rho_{\pi_E})
\end{aligned}
\tag{2.17}
$$

By Gilardoni[27],

$$
D_f(\rho_{\pi_E}, \rho_\pi) \geq \frac{f''(1)}{2} D_{TV}^2(\rho_{\pi_E}, \rho_\pi) \tag{2.18}
$$

Thus we have

$$
V_{\pi_E}-V_\pi \leq H\sqrt{\frac{2}{f''(1)}\epsilon} \tag{2.19}
$$

□

Remark. Theorem 5 is a generalization of corollary 1 in Xu et al.[67]. The distance they considered: KL, Reverse KL, χ^2, JS and Hellinger distance are all special cases of f in theorem 5.

For f-GAIL[72] with the objective:

$$
\min_\pi \max_{f^* \in \mathcal{F}^*, D \in \mathcal{D}} \mathbb{E}_{\pi_E}[D(s,a)]-\mathbb{E}_\pi\left[f^*(D(s,a))\right]-\mathcal{H}(\pi) \tag{2.20}
$$

Similarly, the following holds for f-GAIL,

corollary 6. *Given an expert policy π_E and an imitated policy π, s.t. $\forall f \in \mathcal{F}, D_f(\rho_\pi, \rho_{\pi_E}) \leq \epsilon$(which can be achieved by f-GAIL), where $\mathcal{F}$ is a class of convex functions, and $\forall f \in \mathcal{F}$ is differentiable up to order 3 at $x = 1$ and $f''(1) > 0$. Let $F = \arg\max_{f \in \mathcal{F}} f''(1)$, then we have $V_{\pi_E} - V_\pi \leq H\sqrt{\frac{2}{F}\epsilon}$*

Remark. Above inequalities between f-divergences are not sharp, the sharp ones can be found in Guntuboyina et al.[30] which generally does not have an analytical form, but can be achieved by solving some convex optimization problems.

Above error bounds results show BC has a quadratic dependency on planning horizon while GAIL and its variations have linear dependency. But note that this does not mean GAIL is guaranteed to be superior over BC or f-GAIL is better than GAIL.

2.4.2 Generalization Ability of GAIL

GAIL aims to minimize the discrepancy between the expert and the learner's occupancy measure, however, in practice, we only have finite samples of expert trajectories, drawn from the real distribution. Generalization bounds are introduced to control the exact loss when we can only minimize its empirical version. For analyzing the generalization ability of GAIL, Xu et al.[67] view the imitated policy as to minimize the following neural network distance:

Definition 2.4.1 (Neural Network Distance)**.**

$$d_{\mathcal{D}}(\mu, \nu) = \sup_{D \in \mathcal{D}} \left\{ \mathbb{E}_{(s,a)\sim\mu}[D(s,a)] - \mathbb{E}_{(s,a)\sim\nu}[D(s,a)] \right\} \tag{2.21}$$

Neural Network Distance is also known as Integral Probability Metrics(IPM), which is the objective for Generative Adversarial Networks(GANs), e.g Wasserstein GAN[3]. For the original GAIL and f-divergence related GAIL variants, IPM is not the direct objective of these algorithms. Thus, different from Xu et al.[67], we consider the following neural f distance[71], as in Arora et al.[4], which is a more direct objective for GAIL and its variants:

Definition 2.4.2 (Neural f-Divergence)**.**

$$d_{f,\mathcal{D}}(\mu, \nu) = \sup_{D \in \mathcal{D}} \left\{ \mathbb{E}_{(s,a)\sim\mu}[D(s,a)] - \mathbb{E}_{(s,a)\sim\nu}[D(s,a)] - \mathbb{E}_{(s,a)\sim\mu}\left[f^*(D(s,a))\right] \right\} \tag{2.22}$$

As in Xu et al.[67], we adapt the standard Rademacher complexity to establish generalization bounds:

Definition 2.4.3 (Rademacher Complexity)**.**

$$R_{\mu}^{(m)}(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \tau_i f\left(X_i\right)\right] \tag{2.23}$$

where $X_i \sim \mu, i = 1, ..., m$*,* m *is the number of samples, and random variable* τ_i *:* $\mathrm{P}(\tau_i = 1) = \mathrm{P}(\tau_i = -1) = \frac{1}{2}$

We consider f-VIM, which is a generalized version of GAIL

Theorem 7. *Consider a discriminator class set* $\mathcal{D}$ *with* $\forall(s,a) \in, D \in \mathcal{D}, |D(s,a)| \leq \Delta$*. Given an expert policy* π_E *and an imitated policy* π *with* $d_{f,\mathcal{D}}\left(\hat{\rho}_{\pi_{\mathrm{E}}}, \hat{\rho}_{\pi}\right) - \inf_{\pi\in\Pi} d_{f,\mathcal{D}}\left(\hat{\rho}_{\pi_{\mathrm{E}}}, \hat{\rho}_{\pi}\right) \leq \epsilon$*(which can be achieved by f-VIM), then* $\forall\delta \in (0,1)$*, with probability at least* $1-\delta$*, we have*

$$d_{\mathcal{D}}\left(\rho_{\pi_{\mathrm{E}}}, \rho_{\pi}\right) - d_{\mathcal{D}}\left(\hat{\rho}_{\pi_{\mathrm{E}}}, \hat{\rho}_{\pi}\right) \leq 2\mathcal{R}_{\rho_{\pi_{\mathrm{E}}}}^{(m)}(\mathcal{D}) + 2\mathcal{R}_{\rho_{\pi}}^{(m)}(\mathcal{D}) + 2\mathcal{R}_{\rho_{\pi}}^{(m)}(f^*(\mathcal{D})) + 6\Delta\sqrt{\frac{\log(4/\delta)}{2m}} \tag{2.24}$$

Proof.

$$\begin{aligned}
& d_{\mathcal{D}}\left(\rho_{\pi_{\mathrm{E}}}, \rho_{\pi}\right) - d_{\mathcal{D}}\left(\hat{\rho}_{\pi_{\mathrm{E}}}, \hat{\rho}_{\pi}\right) \\
=& \sup_{D\in\mathcal{D}} \left[\mathbb{E}_{(s,a)\sim\rho_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\rho_{\pi}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\rho_{\pi}}[f^*(D(s,a))]\right] - \\
& \sup_{D\in\mathcal{D}} \left[\mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[f^*(D(s,a))]\right] \\
\leq& \sup_{D\in\mathcal{D}} \{\left[\mathbb{E}_{(s,a)\sim\rho_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\rho_{\pi}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\rho_{\pi}}[f^*(D(s,a))]\right] - \\
& \left[\mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[f^*(D(s,a))]\right]\} \\
\leq& \sup_{D\in\mathcal{D}} \left[\mathbb{E}_{(s,a)\sim\rho_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi_{\mathrm{E}}}}[D(s,a)]\right] + \\
& \sup_{D\in\mathcal{D}} \left[\mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\rho_{\pi}}[D(s,a)]\right] + \\
& \sup_{D\in\mathcal{D}} \left[\mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[f^*(D(s,a))] - \mathbb{E}_{(s,a)\sim\rho_{\pi}}[f^*(D(s,a))]\right] \\
\leq& \sup_{D\in\mathcal{D}} \left|\mathbb{E}_{(s,a)\sim\rho_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi_{\mathrm{E}}}}[D(s,a)]\right| + \\
& \sup_{D\in\mathcal{D}} \left|\mathbb{E}_{(s,a)\sim\rho_{\pi}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[D(s,a)]\right| + \\
& \sup_{D\in f^*(\mathcal{D})} \left|\mathbb{E}_{(s,a)\sim\rho_{\pi}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi}}[D(s,a)]\right|
\end{aligned} \tag{2.25}$$

By McDiarmid's inequality[54],

$$\begin{aligned}
&\sup_{D\in\mathcal{D}} \left|\mathbb{E}_{(s,a)\sim\rho_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi_{\mathrm{E}}}}[D(s,a)]\right| \\
\leq &\mathbb{E}\left[\sup_{D\in\mathcal{D}} \left|\mathbb{E}_{(s,a)\sim\rho_{\pi_{\mathrm{E}}}}[D(s,a)] - \mathbb{E}_{(s,a)\sim\hat{\rho}_{\pi_{\mathrm{E}}}}[D(s,a)]\right|\right] + 2\Delta\sqrt{\frac{\log(4/\delta)}{2m}} \\
\leq &2\mathbb{E}_{\boldsymbol{\sigma},\rho_{\pi_{\mathrm{E}}}}\left[\sup_{D\in\mathcal{D}} \sum_{i=1}^{m} \frac{1}{m}\sigma_i D\left(s^{(i)}, a^{(i)}\right)\right] + 2\Delta\sqrt{\frac{\log(4/\delta)}{2m}} \\
&= 2\mathcal{R}^{(m)}_{\rho_{\pi_{\mathrm{E}}}}(\mathcal{D}) + 2\Delta\sqrt{\frac{\log(4/\delta)}{2m}}
\end{aligned} \tag{2.26}$$

Combining with 2.25, we have

$$d_{\mathcal{D}}\left(\rho_{\pi_{\mathrm{E}}}, \rho_{\pi}\right) - d_{\mathcal{D}}\left(\hat{\rho}_{\pi_{\mathrm{E}}}, \hat{\rho}_{\pi}\right) \leq 2\mathcal{R}^{(m)}_{\rho_{\pi_{\mathrm{E}}}}(\mathcal{D}) + 2\mathcal{R}^{(m)}_{\rho_{\pi}}(\mathcal{D}) + 2\mathcal{R}^{(m)}_{\rho_{\pi}}(f^*(\mathcal{D})) + 6\Delta\sqrt{\frac{\log(4/\delta)}{2m}} \tag{2.27}$$

□

Chapter 3

Conclusion and Future Works

3.1 Summary

In chapter 1, we extend the canonical epidemiological model SIR to find an optimal decision making with the aim to balance between economy and people's health. In our model, people in different health statuses take different decisions on their working hours and consumption to maximise their own utility, while policymakers control the lockdown rate to maximise the overall welfare, which leads to a two phases optimisation problem. Several parameters in our model are not straightforward to specify using the common epidemic data for modelling. We develop a novel method of parameter estimation through various additional sources of data. Our results show that lockdown measures could effectively reduce the deaths and infections caused by the COVID-19. There is an inevitable trade-off between the short-term recession, and health problems caused by the pandemic, and how policymakers deal with this could lead to very different decisions. We quantify the trade-off by emphasising the cost of death in the model objective, which enables the optimal lockdown policy to discover a balance between the economic and epidemic outcomes. The timing of starting and ending the lockdown control policy makes much difference in terms of both the economic and epidemic outcomes. So the earlier to start the control, the better the results will be. It is crucial to avoid premature ending of the control. In the analysis of the smart containment policy, the results suggest that additional information about the health status of people is beneficial, as the optimal lockdown control policy will reach much better outcomes if it could be implemented on people with different health status separately. Through comparison of lockdown policies, we suggest that R_0 cannot be the only foundation for policy-making.

In chapter 2, we summarized imitation learning methods, mainly behavioural cloning, inverse reinforcement learning, generative adversarial imitation learning,

their variants and current applications of imitation learning methods in finance. Although there hasn't been much of literature on this topic, with the increasing availability of financial data in recent years, as a data-driven approach that has shown success in learning from human demonstrations in fields such as robotics and automated driving, imitation learning methods might have more capability of modelling and inference of markets and investors' behaviours comparing to traditional methods.

We also analyse the error bounds and generalization ability of BC, GAIL and its variations.

3.2 Future Works

In Chapter 1, the two-stage optimization problem 1.19 and 1.23 is a bilevel optimization problem[13] and we solved this problem by its KKT condition. However, we acknowledge that the KKT condition is necessary but not sufficient. Therefore, it remains an open question regarding the uniqueness of the equilibrium solution for our specific problem. To further advance the understanding and resolution of this problem, we recognize the need to explore and evaluate alternative solution techniques specifically tailored for bilevel optimization problems. By delving deeper into the literature and actively engaging with the existing methodologies, we aim to broaden our knowledge and identify innovative approaches that could enhance the quality of the obtained solutions. We intend to conduct an extensive study of the existing techniques, examine their applicability, and adapt them to our specific problem domain, thus contributing to the advancement of the overall field of bilevel optimization.

In Chapter 2, for future research, we will study on imitation learning algorithms that are adaptive to applications on financial markets. Below lists some possible topics:

1. Bounded Rationality of Agents. In most of current applications of imitation learning in financial markets, agents(both experts and learners) are assumed to be perfectly rational, but real-world investors are risk-averse and the utility function does not have a definite formation. Ortega et al.[47][48] provide a computation framework for bounded rationality, in which they penalize the reward function by information cost to deviate from the perfect rational behaviour. This is one possible way of generalizing current methods and provide more insight on investors and markets behaviours.

2. Time-Varying Polices. Imitation learning algorithms we discussed in this work are all time-invariant, but in real-world financial markets, behaviours or risk preference of people may change over time. In this case, online-learning algorithms or Bayesian type approaches that could learn polices that can adaptive with time may be a better fit for modelling markets and investors behaviours.

3. Model-Based Methods. Current imitation learning applications on financial markets are mostly model-free approaches that assume a general MDP environment without any prior knowledge. Considering there are various well-established models for markets and investors behaviours, we could try to add model-based information to guide the purely model-free methods.

4. Alternative Generative Algorithms. Recently, generative algorithms, for example diffusion model[62] and DDPM[35] have shown promising performance in image generation and are superior over GANs in many cases. Unlike GANs and generative algorithms in imitation learning, diffusion models do not need adversarial training, so that they are expected to be easier to train and bring more stable results. Pearce et al. [51] made the first attempt to combine diffusion models with Behavioural Cloning to imitate agents behaviours in various environments. Inspiring by this, combing GAIL or other imitation learning algorithms with DDPM may be able to achieve better performance.

Bibliography

[1] Daron Acemoglu, Victor Chernozhukov, Iván Werning, and Michael D Whinston. Optimal targeted lockdowns in a multi-group sir model. *NBER Working Paper*, 27102, 2020.

[2] Fernando E Alvarez, David Argente, and Francesco Lippi. A simple planning problem for covid-19 lockdown. Technical report, National Bureau of Economic Research, 2020.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. arxiv 2017. *arXiv preprint arXiv:1701.07875*, 30(4), 2017.

[4] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232. PMLR, 2017.

[5] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995.

[6] David W Berger, Kyle F Herkenhoff, and Simon Mongey. An seir infectious disease model with testing and conditional quarantine. Technical report, National Bureau of Economic Research, 2020.

[7] Aude G Billard, Sylvain Calinon, and Rüdiger Dillmann. Learning from humans. *Springer handbook of robotics*, pages 1995–2014, 2016.

[8] Kenneth Bogert, Jonathan Feng-Shun Lin, Prashant Doshi, and Dana Kulic. Expectation-maximization for inverse reinforcement learning with hidden data. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1034–1042, 2016.

[9] Luiz Brotherhood, Philipp Kircher, Cezar Santos, and Michèle Tertilt. An economic model of the covid-19 epidemic: The importance of testing and age-specific policies. 2020.

[10] Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

[11] Jae-Deug Choi and Kee-Eung Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12:691–730, 2011.

[12] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4693–4700. IEEE, 2018.

[13] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153:235–256, 2007.

[14] Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine learning in Finance*, volume 1170. Springer, 2020.

[15] Martin S Eichenbaum, Sergio Rebelo, and Mathias Trabandt. The macroeconomics of epidemics. Technical report, National Bureau of Economic Research, 2020.

[16] Maryam Farboodi, Gregor Jarosch, and Robert Shimer. Internal and external effects of social distancing in a pandemic. Technical report, National Bureau of Economic Research, 2020.

[17] Miguel Faria-e Castro. Fiscal policy during a pandemic. *FRB St. Louis Working Paper*, (2020-006), 2020.

[18] Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437(7056):209–214, 2005.

[19] Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, et al. Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. 2020. *DOI*, 10:77482, 2020.

[20] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58. PMLR, 2016.

[21] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

[22] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*.

[23] Bent Fuglede and Flemming Topsoe. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[24] Tanmay Gangwani, Joel Lehman, Qiang Liu, and Jian Peng. Learning belief representations for imitation learning in pomdps. In *Uncertainty in Artificial Intelligence*, pages 1061–1071. PMLR, 2020.

[25] Seyed Kamyar Seyed Ghasemipour, Shane Gu, and Richard Zemel. Understanding the relation between maximum-entropy inverse reinforcement learning and behaviour cloning. 2019.

[26] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pages 1259–1277. PMLR, 2020.

[27] GL Gilardoni. On pinsker's type inequalities and csiszár's f-divergence. *Part I: Second and Fourth-order inequalities (Preprint arXiv: cs/0603097v2)*, 2008.

[28] Martin Gonzalez-Eiras and Dirk Niepelt. On the optimal" lockdown" during an epidemic. Technical report, CESifo Working Paper, 2020.

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[30] Adityanand Guntuboyina, Sujayam Saha, and Geoffrey Schiebinger. Sharp inequalities for f-divergences. *IEEE transactions on information theory*, 60(1):104–121, 2013.

[31] Josiah Hanna and Peter Stone. Grounded action transformation for robot learning in simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[32] Tiberiu Harko, Francisco SN Lobo, and MK Mak. 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, 2014.

[33] Herbert W Hethcote. Three basic epidemiological models. In *Applied mathematical ecology*, pages 119–144. Springer, 1989.

[34] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[36] Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 25(2):328–373, 2013.

[37] Callum J Jones, Thomas Philippon, and Venky Venkateswaran. Optimal mitigation policies in a pandemic: Social distancing and working from home. Technical report, National Bureau of Economic Research, 2020.

[38] Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*, pages 313–329. Springer, 2021.

[39] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[40] Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pages 5496–5505. PMLR, 2021.

[41] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*, 2020.

[42] Jonathan Lacotte, Mohammad Ghavamzadeh, Yinlam Chow, and Marco Pavone. Risk-sensitive generative adversarial imitation learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2154–2163. PMLR, 2019.

[43] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. *Advances in Neural Information Processing Systems*, 30, 2017.

[44] Robert Marschinski, Pietro Rossi, Massimo Tavoni, and Flavio Cocco. Portfolio selection with probabilistic utility. *Annals of Operations Research*, 151(1):223–239, 2007.

[45] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

[46] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[47] Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013.

[48] Pedro A Ortega, Daniel A Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.

[49] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

[50] Brahma S Pavse, Faraz Torabi, Josiah Hanna, Garrett Warnell, and Peter Stone. Ridm: Reinforced inverse dynamics modeling for learning from a single observed demonstration. *IEEE Robotics and Automation Letters*, 5(4):6262–6269, 2020.

[51] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023.

[52] Facundo Piguillem and Liyan Shi. Optimal covid-19 quarantine and testing policies. 2020.

[53] Qifeng Qiao and Peter A Beling. Inverse reinforcement learning with gaussian process. In *Proceedings of the 2011 American control conference*, pages 113–118. IEEE, 2011.

[54] Emmanuel Rio. On mcdiarmid's concentration inequality. 2013.

[55] Jacobo Roa-Vicens, Cyrine Chtourou, Angelos Filos, Francisco Rullan, Yarin Gal, and Ricardo Silva. Towards inverse reinforcement learning for limit order book dynamics. *arXiv preprint arXiv:1906.04813*, 2019.

[56] Jacobo Roa-Vicens, Yuanbo Wang, Virgile Mison, Yarin Gal, and Ricardo Silva. Adversarial recovery of agent rewards from latent spaces of the limit order book. *arXiv preprint arXiv:1912.04242*, 2019.

[57] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[58] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[59] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[60] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[61] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[62] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[63] Umar Syed and Robert E Schapire. A reduction from apprenticeship learning to classification. *Advances in neural information processing systems*, 23, 2010.

[64] Huaiyu Tian, Yonghong Liu, Yidan Li, Chieh-Hsi Wu, Bin Chen, Moritz UG Kraemer, Bingying Li, Jun Cai, Bo Xu, Qiqi Yang, et al. An investigation of transmission control measures during the first 50 days of the covid-19 epidemic in china. *Science*, 368(6491):638–642, 2020.

[65] Haoran Wang and Shi Yu. Robo-advising: Enhancing investment with inverse optimization and deep reinforcement learning. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 365–372. IEEE, 2021.

[66] Ning Wang, Yuting Fu, Hu Zhang, and Huipeng Shi. An evaluation of mathematical models for the outbreak of covid-19. *Precision Clinical Medicine*, 2020.

[67] Tian Xu, Ziniu Li, and Yang Yu. Error bounds of imitating policies and environments. *Advances in Neural Information Processing Systems*, 33:15737–15749, 2020.

[68] Steve Y Yang, Qifeng Qiao, Peter A Beling, William T Scherer, and Andrei A Kirilenko. Gaussian process-based algorithmic trading strategy identification. *Quantitative Finance*, 15(10):1683–1703, 2015.

[69] Steve Y Yang, Yangyang Yu, and Saud Almahdi. An investor sentiment reward-based trading system using gaussian inverse reinforcement learning algorithm. *Expert Systems with Applications*, 114:388–401, 2018.

[70] Shi Yu, Haoran Wang, and Chaosheng Dong. Learning risk preferences from investment portfolios using inverse optimization. *Research in International Business and Finance*, page 101879, 2023.

[71] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

[72] Xin Zhang, Yanhua Li, Ziming Zhang, and Zhi-Li Zhang. f-gail: Learning f-divergence for generative adversarial imitation learning. *Advances in neural information processing systems*, 33:12805–12815, 2020.

[73] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

[74] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.